SIAM/ASA J. Uncertainty Quantification
Vol. 5, pp. 1136–1165

Gaussian Approximations for Probability Measures on $\mathbb{R}^{d*}$

Yulong Lu[†], Andrew Stuart[‡], and Hendrik Weber[†]

Abstract. This paper concerns the approximation of probability measures on $\mathbb{R}^d$ with respect to the Kullback–Leibler divergence. Given an admissible target measure, we show the existence of the best approximation, with respect to this divergence, from certain sets of Gaussian measures and Gaussian mixtures. The asymptotic behavior of such best approximations is then studied in the small parameter limit where the measure concentrates; this asympotic behavior is characterized using Γ-convergence. The theory developed is then applied to understand the frequentist consistency of Bayesian inverse problems in finite dimensions. For a fixed realization of additive observational noise, we show the asymptotic normality of the posterior measure in the small noise limit. Taking into account the randomness of the noise, we prove a Bernstein–Von Mises type result for the posterior measure.

Key words. Gaussian approximation, Kullback–Leibler divergence, Gamma-convergence, Bernstein–Von Mises theorem

AMS subject classifications. 60B10, 60H07, 62F15

DOI. 10.1137/16M1105384

1. Introduction.

In this paper, we study the "best" approximation of a general finite dimensional probability measure, which could be non-Gaussian, from a set of simple probability measures, such as a single Gaussian measure or a Gaussian mixture family. We define "best" to mean the measure within the simple class which minimizes the Kullback–Leibler divergence between itself and the target measure. This type of approximation is central to many ideas, especially including the so-called variational inference [30], that are widely used in machine learning [3]. Yet such approximation has not been the subject of any substantial systematic underpinning theory. The purpose of this paper is to develop such a theory in the concrete finite dimensional setting in two ways: (i) by establishing the existence of best approximations and (ii) by studying their asymptotic properties in a measure concentration limit of interest. The abstract theory is then applied to study frequentist consistency [28] of Bayesian inverse problems.

1.1. Background and overview.

The idea of approximation for probability measures with respect to Kullback–Leibler divergence has been applied in a number of areas; see, for

*Received by the editors November 28, 2016; accepted for publication (in revised form) June 22, 2017; published electronically November 16, 2017.

http://www.siam.org/journals/juq/5/M110538.html

Funding: The first author was supported by EPSRC as part of MASDOC DTC at the University of Warwick with grant EP/HO23364/1. The second author was supported by DARPA, EPSRC, and ONR. The third author was supported by the Royal Society through University Research Fellowship UF140187.

[†]Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK (yulong.lu@warwick.ac.uk, hendrik.weber@warwick.ac.uk).

[‡]Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 (astuart@caltech.edu).

example, [20, 15, 19, 24]. Despite the wide usage of Kullback–Leibler approximation, systematic theoretical study has only been initiated recently. In [23], the measure approximation problem is studied from the calculus of variations point of view and existence of minimizers established therein; the companion paper [22] proposed numerical algorithms for implementing Kullback–Leibler minimization in practice. In [19], Gaussian approximation is used as a new approach for identifying the most likely path between equilibrium states in molecular dynamics; furthermore, the asymptotic behavior of the Gaussian approximation in the small temperature limit is analyzed via Γ-convergence. Here our interest is to develop the ideas in [19] in the context of a general class of measure approximation problems in finite dimensions.

To be concrete we consider approximation of a family of probability measures $\{\mu_\varepsilon\}_{\varepsilon>0}$ on $\mathbb{R}^d$ with (Lebesgue) density of the form

$$\mu_\varepsilon(dx) = \frac{1}{Z_{\mu,\varepsilon}} \exp\left(-\frac{1}{\varepsilon}V_1^\varepsilon(x) - V_2(x)\right) dx; \tag{1}$$

here $Z_{\mu,\varepsilon}$ is the normalization constant. A typical example of a measure μ_ε with this form is a posterior measure in Bayesian inverse problems. For instance, consider the inverse problem of identifying x from a sequence of noisy observations $\{y_j\}_{j\in\mathbb{N}}$, where

$$y_j = G(x) + \eta_j,$$

and where the η_j denote the random noise terms. This may model a statistical measurement with an increasing number of observations or with vanishing noise. In the Bayesian approach to this inverse problem, if we take a prior with density proportional to $\exp(-V_2(x))$, then the posterior measure is given by (1) with the function $\varepsilon^{-1}V_1^\varepsilon$, up to an additive constant, coinciding with the negative log-likelihood. The parameter ε is associated with the number of observations or the noise level of the statistical experiment.

Our study of Gaussian approximation to the measures μ_ε in (1) is partially motivated by the famous Bernstein–Von Mises (BvM) theorem [28] in asymptotic statistics. Roughly speaking, the BvM theorem states that under mild conditions on the prior, the posterior distribution of a Bayesian procedure converges to a Gaussian distribution centered at any consistent estimator (for instance, the maximum likelihood estimator (MLE)) in the limit of large data (or, relatedly, small noise [5]). The BvM theorem is of great importance in Bayesian statistics for at least two reasons. First, it gives a quantitative description of how the posterior contracts to the underlying truth. Second, it implies that the Bayesian credible sets are asymptotically equivalent to frequentist confidence intervals and hence the estimation of the latter can be realized by making use of the computational power of Markov chain Monte Carlo algorithms. We interpret the BvM phenomenon in the abstract theoretical framework of best Gaussian approximations with respect to a Kullback–Leibler measure of divergence.

1.2. Main contributions.

The main contributions of this paper are twofold:

- We use the calculus of variations to give a framework to the problem of finding the best Gaussian (mixture) approximation of a given measure, with respect to a Kullback–Liebler divergence;
- We study the resulting calculus of variations problem in the small noise (or large data) limits, therby making new links to, and ways to think about, the classical BvM theory of asymptotic normality.

We describe these contributions in more detail. First we introduce a theoretical framework of calculus of variations to analyze the measure approximation problem. Given a measure μ_ε defined by (1), we find a measure ν_ε from a set of simple measures, Gaussians, or mixtures of finitely many Gaussians, which minimizes the Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$. We characterize the limiting behavior of the best approximation ν_ε as well as the limiting behavior of the Kullback–Leibler divergence as $\varepsilon \downarrow 0$ using the framework of Γ-convergence. In particular, if μ_ε is a multimodal distribution and ν_ε is the best approximation from within the class of Gaussian mixtures, then the limit of the minimized Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon)$ can be characterized explicitly as the sum of two contributions: a local term which consists of a weighted sum of the Kullback–Leibler divergences between the Gaussian approximations, as well as the Gaussian measure whose covariance is determined by the Hessian of V_2 at its minimizers, and a global term which measures how well the weights approximate the mass distribution between the modes; see Theorem 4.2.

We then adopt the abstract measure approximation theory to understanding the posterior consistency of finite dimensional Bayesian inverse problems. In particular, we give an alternative (and more analytical) proof of the BvM theorem; see Theorem 5.4 and Corollary 5.5. We highlight the fact that our BvM result improves classical BvM results for parametric statistical models in two aspects. First, the convergence of posterior in the total variation distance is improved to convergence in the Kullback–Leibler divergence, under certain regularity assumptions on the forward map. Second, our BvM result allows the posterior distribution to be multimodal, in which case the posterior approaches a mixture of Gaussian distributions rather than a single Gaussian distribution in the limit of infinite data. These improvements come at a cost, and we need to make stronger assumptions than those made in classical BvM theory.

1.3. Structure. The rest of the paper is organized as follows. In section 2 we set up various underpinning concepts which are used throughout the paper: in subsections 2.1 and 2.2, we recall some basic facts on Kullback–Leibler divergence and Γ-convergence and in subsections 2.3 and 2.4 we spell out the assumptions made and the notation used. In sections 3 and 4 we study the problem of approximation of the measure μ_ε by, respectively, a single Gaussian measure and a Gaussian mixture. In particular, the small ε asymptotics of the Gaussians (or Gaussian mixtures) are captured by using the framework of Γ-convergence. In section 5, the theory which we have developed is applied to understand the posterior consistency for Bayesian inverse problems, and connections to the BvM theory. Finally, we finish in section 6 with several conclusion remarks.

2. Set-up.

2.1. Kullback–Leibler divergence. Let ν and μ be two probability measures on $\mathbb{R}^d$ and assume that ν is absolutely continuous with respect to μ. The Kullback–Leibler divergence, or relative entropy, of ν with respect to μ is

$$D_{\mathrm{KL}}(\nu||\mu) = \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu}\right).$$

If ν is not absolutely continuous with respect to μ, then the Kullback–Leibler divergence is defined as $+\infty$. By definition, the Kullback–Leibler divergence is nonnegative but it is not a metric since it does not obey the triangle inequality and it is not symmetric in its two arguments. In this paper, we will consider minimizing $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$ with respect to ν, over a suitably chosen set of measures, and with μ_ε being the target measure defined in (1). Swapping the order of these two measures within the divergence is undesirable for our purposes. This is because minimizing $D_{\mathrm{KL}}(\mu_\varepsilon||\cdot)$ within the set of all Gaussian measures will lead to matching of moments [3]; this is inappropriate for multimodal measures where a more desirable outcome would be the existence of multiple local minimizers at each mode [23, 22].

Although the Kullback–Leibler divergence is not a metric, its information theoretic interpretation make it natural for approximate inference. Furthermore it is a convenient quantity to work with for at least two reasons. First, the divergence provides a useful upper bound for many metrics; in particular, one has the Pinsker inequality

$$d_{\mathrm{TV}}(\nu,\mu) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\nu||\mu)}, \tag{2}$$

where d_{TV} denotes the total variation distance. Second, the logarithmic structure of $D_{\mathrm{KL}}(\cdot||\cdot)$ allows us to carry out explicit calculations, and numerical computations, which are considerably more difficult when using the total variation distance directly.

2.2. Γ-convergence. We recall the definition and a basic result concerning Γ-convergence. This is a useful tool for studying families of minimization problems. In this paper we will use it to study the parametric limit $\varepsilon \to 0$ in our approximation problem.

Definition 2.1. *Let $\mathcal{X}$ be a metric space and $E_\varepsilon : \mathcal{X} \to \mathbb{R}$ a family of functionals indexed by $\varepsilon > 0$. Then E_ε Γ-converges to $E : \mathcal{X} \to \mathbb{R}$ as $\varepsilon \to 0$ if the following conditions hold:*

(i) *(liminf inequality) for every $u \in \mathcal{X}$, and for every sequence $u_\varepsilon \in \mathcal{X}$ such that $u_\varepsilon \to u$, it holds that $E(u) \leq \liminf_{\varepsilon \downarrow 0} E_\varepsilon(u_\varepsilon)$;*

(ii) *(limsup inequality) for every $u \in \mathcal{X}$ there exists a recovery sequence $\{u_\varepsilon\}$ such that $u_\varepsilon \to u$ and $E(u) \geq \limsup_{\varepsilon \downarrow 0} E_\varepsilon(u_\varepsilon)$.*

We say a sequence of functionals $\{E_\varepsilon\}$ is compact if $\limsup_{\varepsilon \downarrow 0} E_\varepsilon(u_\varepsilon) < \infty$ implies that there exists a subsequence $\{u_{\varepsilon_j}\}$ such that $u_{\varepsilon_j} \to u \in \mathcal{X}$.

The notion of Γ-convergence is useful because of the following fundamental theorem, which can be proved by similar methods as in the proof of [4, Theorem 1.21].

Theorem 2.2. *Let u_ε be a minimizer of E_ε with $\limsup_{\varepsilon \downarrow 0} E_\varepsilon(u_\varepsilon) < \infty$. If E_ε is compact and Γ-converges to E, then there exists a subsequence u_{ε_j} such that $u_{\varepsilon_j} \to u$, where u is a minimizer of E.*

Thus, when this theorem applies, it tells us that minimizers of E characterize the limits of convergent subsequences of minimizers of E_ε. In other words the Γ-limit captures the behavior of the minimization problem in the small ε limit.

2.3. Assumptions. Throughout the paper, we make the following assumptions on the potential functions V_1^ε and V_2 which define the target measure of interest.

Assumption 2.3.

(A-1) *For any* $\varepsilon > 0$, V_1^ε *and* V_2 *are nonnegative functions in the space* $C^4(\mathbb{R}^d)$ *and* $C^2(\mathbb{R}^d)$, *respectively. Moreover, there exists constants* $\varepsilon_0 > 0$ *and* $M_V > 0$ *such that when* $\varepsilon < \varepsilon_0$,

$$|\partial_x^\alpha V_1^\varepsilon(x)| \vee |\partial_x^\beta V_2(x)| \le M_V e^{|x|^2}$$

for any $|\alpha| \le 4, |\beta| \le 2$ *and all* $x \in \mathbb{R}^d$.

(A-2) *There exists* $n > 0$ *such that when* $\varepsilon \ll 1$, *the set of minimizers of* V_1^ε *is* $\mathscr{E}^\varepsilon = \{x_\varepsilon^1, x_\varepsilon^2, \dots, x_\varepsilon^n\}$ *and* $V_1^\varepsilon(x_\varepsilon^i) = 0, i = 1, \dots, n$.

(A-3) *There exists* V_1 *such that* $V_1^\varepsilon \to V_1$ *pointwise. The limit* V_1 *has* n *distinct global minimizers which are given by* $\mathscr{E} = \{x^1, x^2, \dots, x^n\}$. *For each* $i = 1, \dots, n$ *the Hessian* $D^2V_1(x^i)$ *is positive definite.*

(A-4) *The convergence* $x_\varepsilon^i \to x^i$ *holds.*

(A-5) *There exist constants* $c_0, c_1 > 0$ *and* $\varepsilon_0 > 0$ *such that when* $\varepsilon < \varepsilon_0$,

$$V_1^\varepsilon(x) \ge -c_0 + c_1|x|^2, x \in \mathbb{R}^d.$$

Remark 2.4. Conditions (A-2)–(A-4) mean that for sufficiently small $\varepsilon > 0$, the function V_1^ε behaves like a quadratic function in the neighborhood of the minimizers x_ε^i and of x^i. Consequently, the measure μ_ε is asymptotically normal in the local neighborhood of x_ε^i. In particular, in conjunction with condition (A-5) this implies that there exists $\delta > 0$ and $C_\delta > 0$ such that $\forall\, 0 \le \eta < \delta$,

$$\operatorname{dist}(x, \mathscr{E}) \ge \eta \Longrightarrow \liminf_{\varepsilon \downarrow 0} V_1^\varepsilon(x) \ge C_\delta |\eta|^2. \tag{3}$$

Remark 2.5. The local boundedness of V_1^ε in $C^4(\mathbb{R}^d)$ (assumption (A-1)) together with the pointwise convergence of V_1^ε to V_1 (assumption (A-3)) implies the much stronger locally uniform convergence of derivatives up to order 3. Furthermore, (A-4) then implies that $V_1^\varepsilon(x_\varepsilon^i) \to V_1(x^i)$ and $D^2V_1^\varepsilon(x_\varepsilon^i) \to D^2V_1(x^i)$.

2.4. Notation. Throughout the paper, C and $\tilde{C}$ will be generic constants which are independent of the quantities of interest and may change from line to line. Let $\mathcal{S}_\ge(\mathbb{R}, d)$ and $\mathcal{S}_>(\mathbb{R}, d)$ be the set of all $d \times d$ real matrices which are positive semidefinite or positive definite, respectively. Denote by $N(m, \mathbf{\Sigma})$ a Gaussian measure with mean m and covariance matrix $\mathbf{\Sigma}$. We use $|\mathbf{A}|$ to denote the Frobenius norm of the $d \times d$ matrix $\mathbf{A}$, namely, $|\mathbf{A}| = \sqrt{\operatorname{Tr}(\mathbf{A}^T\mathbf{A})}$. We denote by $\lambda_{\min}(\mathbf{A})$ the smallest eigenvalue of $\mathbf{A}$. We let $B(x, r)$ denote a ball in $\mathbb{R}^d$ with center x and radius r. Given a random variable η, we use $\mathbb{E}^\eta$ and $\mathbb{P}^\eta$ when computing the expectation and the probability under the law of η, respectively.

3. Approximation by single Gaussian measures.

Let $\mathcal{A}$ be the set of Gaussian measures on $\mathbb{R}^d$, given by

$$\mathcal{A} = \{N(m, \mathbf{\Sigma}) : m \in \mathbb{R}^d, \mathbf{\Sigma} \in \mathcal{S}_\ge(\mathbb{R}, d)\}.$$

The set $\mathcal{A}$ is closed with respect to weak convergence of probability measures. Consider the variational problem

$$\inf_{\nu\in\mathcal{A}} D_{\mathrm{KL}}(\nu||\mu_\varepsilon). \tag{4}$$

Given $\nu = N(m,\mathbf{\Sigma}) \in \mathcal{A}$, the Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$ can be calculated explicitly as

$$\begin{aligned} D_{\mathrm{KL}}(\nu||\mu_\varepsilon) &= \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu_\varepsilon}\right) \\ &= \frac{1}{\varepsilon}\mathbb{E}^\nu V_1^\varepsilon(x) + \mathbb{E}^\nu V_2(x) - \log\sqrt{(2\pi)^d \det\mathbf{\Sigma}} - \frac{d}{2} + \log Z_{\mu,\varepsilon}. \end{aligned} \tag{5}$$

If $\mathbf{\Sigma}$ is noninvertible, then $D_{\mathrm{KL}}(\nu||\mu_\varepsilon) = +\infty$. The term $-\frac{d}{2}$ comes from the expectation $\mathbb{E}^\nu \frac{1}{2}(x-m)^T\mathbf{\Sigma}(x-m)$ and is independent of $\mathbf{\Sigma}$. The term $-\log\sqrt{(2\pi)^d\det\mathbf{\Sigma}}$ prevents the measure ν from being too close to a Dirac measure. The following theorem shows that the problem (4) has a solution.

Theorem 3.1. *Consider the measure μ_ε given by (1). For any $\varepsilon > 0$, there exists at least one probability measure $\overline{\nu}_\varepsilon \in \mathcal{A}$ solving the problem (4).*

Proof. We first show that the infimum of (4) is finite. In fact, consider $\nu^* = N(0, \frac{1}{4}\mathbf{I}_d)$. Under Assumption 2.3(A-1) we have that

$$\mathbb{E}^{\nu^*} V_1^\varepsilon(x) \vee \mathbb{E}^{\nu^*} V_2(x) \le \frac{M_V}{\sqrt{(2\pi\times\frac{1}{4})^d}} \int_{\mathbb{R}^d} e^{-\frac{4}{2}|x|^2+|x|^2} dx < \infty.$$

Note that the integral in the last expression is finite due to $-\frac{4}{2}+1<0$. Hence we know from (5) that $\inf_{\nu\in\mathcal{A}} D_{\mathrm{KL}}(\nu||\mu_\varepsilon) < \infty$. Then the existence of minimizers follows from the fact that the Kullback–Leibler divergence has compact sublevel sets and the closedness of $\mathcal{A}$ with respect to weak convergence of probability measures; see, e.g., [23, Corollary 2.2]. ∎

We aim to understand the asymptotic behavior of the minimizers $\overline{\nu}_\varepsilon$ of the problem (4) as $\varepsilon \downarrow 0$. Due to the factor $\frac{1}{\varepsilon}$ in front of V_1^ε in the definition of μ_ε, (1), we expect the typical size of fluctuations around the minimizers to be of order $\sqrt{\varepsilon}$ and we reflect that in our choice of scaling. More precisely, for $m\in\mathbb{R}^d$, $\mathbf{\Sigma}\in\mathcal{S}_{\ge}(\mathbb{R},d)$ we define $\nu_\varepsilon = N(m,\varepsilon\mathbf{\Sigma})$ and set

$$F_\varepsilon(m,\mathbf{\Sigma}) := D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon). \tag{6}$$

Understanding the asymptotic behavior of minimizers $\overline{\nu}_\varepsilon$ in the small ε limit may be achieved by understanding Γ-convergence of the functional F_ε.

To that end, we define weights

$$\beta^i = \left(\det D^2 V_1(x^i)\right)^{-\frac{1}{2}} \cdot e^{-V_2(x^i)}, \qquad i=1,\dots,n,$$

and the counting probability measure on $\{1,\dots,n\}$ given by

$$\boldsymbol{\beta} := \frac{1}{\sum_{j=1}^n \beta^j}(\beta^1,\dots,\beta^n).$$

Intuitively, as $\varepsilon \downarrow 0$, we expect the measure μ_ε to concentrate on the set $\{x^i\}$ with weights on each x^i given by $\boldsymbol{\beta}$; this intuition is reflected in the asymptotic behavior of the normalization constant $Z_{\mu,\varepsilon}$, as we now show. By definition,

$$Z_{\mu,\varepsilon} = \int_{\mathbb{R}^d} \exp\left(-\frac{1}{\varepsilon}V_1^\varepsilon(x) - V_2(x)\right) dx.$$

The following lemma follows from the Laplace approximation for integrals (see, e.g., [14]) and Assumption 2.3(A-4).

Lemma 3.2. *Let V_1^ε and V_2 satisfy Assumption 2.3. Then as $\varepsilon \downarrow 0$,*

$$Z_{\mu,\varepsilon} = \sqrt{(2\pi\varepsilon)^d} \cdot \left(\sum_{i=1}^n \beta^i\right) \cdot (1 + o(1)). \tag{7}$$

Recall from (6) that $F_\varepsilon(m, \boldsymbol{\Sigma}) = D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon)$ with the specific scaling $\nu_\varepsilon = N(m, \varepsilon\boldsymbol{\Sigma})$. In view of the expression (5) for the Kullback–Leibler divergence, it follows from Lemma 3.2 that

$$F_\varepsilon(m, \boldsymbol{\Sigma}) = \frac{1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) - \frac{d}{2} - \frac{1}{2}\log\left(\det \boldsymbol{\Sigma}\right) + \log\left(\sum_{i=1}^n \beta^i\right) + o(1). \tag{8}$$

Armed with this analysis of the normalization constant we may now prove the following theorem which identifies the Γ-limit of F_ε. To this end we define

$$F_0(m, \boldsymbol{\Sigma}) := V_2(m) + \frac{1}{2}\mathrm{Tr}\left(D^2 V_1(m) \cdot \boldsymbol{\Sigma}\right) - \frac{d}{2} - \frac{1}{2}\log\det\boldsymbol{\Sigma} + \log\left(\sum_{i=1}^n \beta^i\right).$$

Theorem 3.3. *The Γ-limit of F_ε is*

$$F(m, \boldsymbol{\Sigma}) := \begin{cases} F_0(m, \boldsymbol{\Sigma}) & \text{if } m \in \mathscr{E} \text{ and } \boldsymbol{\Sigma} \in \mathcal{S}_>(\mathbb{R}, d), \\ \infty & \text{otherwise.} \end{cases} \tag{9}$$

The following corollary follows directly from the Γ-convergence of F_ε.

Corollary 3.4. *Let $\{(m_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)\}$ be a family of minimizers of $\{F_\varepsilon\}$. Then there exists a subsequence $\{\varepsilon_k\}$ such that $(m_{\varepsilon_k}, \boldsymbol{\Sigma}_{\varepsilon_k}) \to (m, \boldsymbol{\Sigma})$ and $F_{\varepsilon_k}(m_{\varepsilon_k}, \boldsymbol{\Sigma}_{\varepsilon_k}) \to F(m, \boldsymbol{\Sigma})$. Moreover, $(m, \boldsymbol{\Sigma})$ is a minimizer of F.*

Before we give the proof of Theorem 3.3, let us first discuss the limit functional F as well as its minimization. We assume that $m = x^{i_0}$ for some $i_0 \in \{1, \dots, n\}$ and rewrite the definition of $F_0(x^{i_0}, \boldsymbol{\Sigma})$ by adding and subtracting $\log(\beta^{i_0}) = -V_2(x^{i_0}) - \frac{1}{2}\log\left(\left(\det D^2 V_1(x^{i_0})\right)\right)$ and cancelling the terms involving $V_2(x^{i_0})$ as

$$\begin{aligned} F_0(x^{i_0}, \boldsymbol{\Sigma}) = {} & \frac{1}{2}\mathrm{Tr}\left(D^2 V_1(x^{i_0}) \cdot \boldsymbol{\Sigma}\right) - \frac{d}{2} - \frac{1}{2}\log\det(D^2 V_1(x^{i_0}) \cdot \boldsymbol{\Sigma}) \\ & + \log\left(\sum_{i=1}^n \beta^i\right) - \log\left(\beta^{i_0}\right). \end{aligned} \tag{10}$$

Now it is interesting to see that the first line of (10) gives the Kullback–Leibler divergence $D_{\mathrm{KL}}\left(N(x^{i_0},\mathbf{\Sigma}) \,||\, N(x^{i_0},(D^2V_1(x^{i_0}))^{-1})\right)$. The second line of (10) is equal to the Kullback–Leibler divergence $D_{\mathrm{KL}}(\mathbf{e^{i_0}} \,||\, \boldsymbol{\beta})$, for $\mathbf{e^{i_0}} := (\mathbf{0},\dots,\mathbf{1},\dots,\mathbf{0})$. In conclusion,

$$F_0(x^i,\mathbf{\Sigma}) = D_{\mathrm{KL}}\left(N(x^i,\mathbf{\Sigma}) \,||\, N(x^i,(D^2V_1(x^i))^{-1})\right) + D_{\mathrm{KL}}(\mathbf{e^i} \,||\, \boldsymbol{\beta}). \tag{11}$$

In other words, in the limit $\varepsilon \downarrow 0$, the Kullback–Leibler divergence between the best Gaussian measure ν_ε and the measure μ_ε consists of two parts: the first part is the relative entropy between the Gaussian measure with rescaled covariance $\mathbf{\Sigma}$ and the Gaussian measure with covariance determined by $(D^2V_1(x^i))^{-1}$; the second part is the relative entropy between the Dirac mass supported at x^i and a weighted sum of Dirac masses, with weights $\boldsymbol{\beta}$, at the $\{x^j\}_{j=1}^n$. Clearly, to minimize $F_0(m,\mathbf{\Sigma})$, on the one hand, we need to choose $m = x^i$ and $\mathbf{\Sigma} = (D^2V_1(x^i))^{-1}$ for some $i \in 1,\dots,n$; for this choice the first term on the right side of (10) vanishes. In order to minimize the second term we need to choose the minimum x^i with maximal weight β^i. In particular, the following corollary holds.

Corollary 3.5. *The minimum of F_0 is zero when $n = 1$, but it is strictly positive when $n > 1$.*

Corollary 3.5 reflects the fact that, in the limit $\varepsilon \downarrow 0$, a single Gaussian measure is not the best choice for approximating a non-Gaussian measure with multiple modes; this motivates our study of Gaussian mixtures in section 4.

The proofs of Theorem 3.3 and Corollary 3.4 are provided after establishing a sequence of lemmas. The following lemma shows that the sequence of functionals $\{F_\varepsilon\}$ is compact (recall Definition 2.1). It is well known that the Kullback–Leibler divergence (with respect to a fixed reference μ) has compact sublevel sets with respect to weak convergence of probability measures. Here we prove a stronger statement, which is specific to the family of reference measures μ_ε, namely, a uniform bound from above and below for the rescaled covariances, i.e., we prove a bound from above and below for $\mathbf{\Sigma}_\varepsilon$ if we control $F_\varepsilon(m_\varepsilon,\mathbf{\Sigma}_\varepsilon)$.

Lemma 3.6. *Let $\{(m_\varepsilon,\mathbf{\Sigma}_\varepsilon)\} \subset \mathbb{R}^d \times \mathcal{S}_{\geq}(\mathbb{R},d)$ be such that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon,\mathbf{\Sigma}_\varepsilon) < \infty$. Then*

$$0 < \liminf_{\varepsilon\downarrow 0} \lambda_{\min}(\mathbf{\Sigma}_\varepsilon) < \limsup_{\varepsilon\downarrow 0} \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) < \infty \tag{12}$$

and $dist(m_\varepsilon,\mathscr{E}) \downarrow 0$ as $\varepsilon \downarrow 0$. In particular, there exist common subsequences $\{m_k\}_{k\in\mathbb{N}}$ of $\{m_\varepsilon\}$, $\{\mathbf{\Sigma}_k\}_{k\in\mathbb{N}}$ of $\{\mathbf{\Sigma}_\varepsilon\}$ such that $m_k \to x_{i_0}$ with $1 \leq i_0 \leq n$ and $\mathbf{\Sigma}_k \to \mathbf{\Sigma} \in \mathcal{S}_{>}(\mathbb{R},d)$.

Proof. Let $M := \limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon,\mathbf{\Sigma}_\varepsilon) < \infty$. Since m_ε and $\mathbf{\Sigma}_\varepsilon$ are defined in finite dimensional spaces, we only need to show that both sequences are uniformly bounded. The proof consists of the following steps.

Step 1. We first prove the following rough bounds for $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)$: there exist positive constants C_1, C_2 such that when $\varepsilon \ll 1$,

$$C_1 \leq \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) \leq \frac{C_2}{\varepsilon}. \tag{13}$$

In fact, from the formula (8) and the assumption that V_1^ε and V_2 are nonnegative, we can get that when $\varepsilon \ll 1$

$$\log(\det \mathbf{\Sigma}_\varepsilon) \geq 2(C_V - M - 1), \tag{14}$$

where the constant

$$C_V := -\frac{d}{2} + \log\left(\sum_{i=1}^{n} \beta^i\right).$$

Then the lower bound of (13) follows from (14) and the arithmetic-geometric mean inequality

$$\det \mathbf{A} \le \left(\frac{1}{d}\mathrm{Tr}(\mathbf{A})\right)^d, \tag{15}$$

which holds for any positive definite $\mathbf{A}$. In addition, using the condition (A-5) for the potential V_1^ε, we obtain from (8) that when $\varepsilon \ll 1$,

$$\begin{aligned} M &\ge F_\varepsilon(m_\varepsilon, \mathbf{A}_\varepsilon) \\ &\ge \mathbb{E}^{\nu_\varepsilon} V_2(x) + \frac{c_1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon}|x|^2 - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) + C_V - 1 \\ &= \mathbb{E}^{\nu_\varepsilon} V_2(x) + c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) + \frac{c_1|m_\varepsilon|^2}{\varepsilon} - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) + C_V - 1 \\ &\ge c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\left(\frac{1}{d}\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)\right)^d\right) + C_V - 1 \\ &= c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{c_0}{\varepsilon} - \frac{d}{2}\log(\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)) + \frac{d\log d}{2} + C_V - 1, \end{aligned} \tag{16}$$

where we have used the inequality (15) and the assumption that V_2 is nonnegative. Dropping the nonnegative terms on the right-hand side we rewrite this expression as an estimate on $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)$,

$$c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{d}{2}\log(\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)) \le M + \frac{c_0}{\varepsilon} + 1,$$

and conclude that there exists $C_2 > 0$ such that $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) \le C_2/\varepsilon$ by observing that for $x \gg 1$ we have $c_1 x - \frac{d}{2}\log x \ge \frac{c_1}{2}x$.

Step 2. In this step we show that for $\varepsilon \ll 1$ the mass of ν_ε concentrates near the minimizers. More precisely, we claim that there exist constants $R_1, R_2 > 0$, such that for every $\varepsilon \ll 1$ there exists an index $i_0 \in \{1, 2, \ldots, n\}$ such that

$$\nu_\varepsilon\left(B\left(x_{i_0}, \sqrt{\varepsilon(R_1 + R_2 \log\left(\det \mathbf{\Sigma}_\varepsilon\right))}\right)\right) \ge \frac{1}{2n}. \tag{17}$$

On the one hand, from the expression (8) and the assumption that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) \le M$ we know that there exist $C_3, C_4 > 0$ such that when $\varepsilon \ll 1$

$$\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) \le \varepsilon\left(C_3 + C_4 \log\left(\det \mathbf{\Sigma}_\varepsilon\right)\right). \tag{18}$$

On the other hand, it follows from (3) that for $\eta \ll 1$

$$\begin{aligned} \mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) &\ge \mathbb{E}^{\nu_\varepsilon}\left[V_1^\varepsilon(x)\mathbf{I}_{(\cup_{i=1}^n B(x_i,\eta))^c}(x)\right] \\ &\ge C_\delta \eta^2 \nu_\varepsilon(\cup_{i=1}^n B(x_i,\eta))^c, \end{aligned} \tag{19}$$

which combined with (18) leads to

$$\nu_\varepsilon(\cup_{i=1}^n B(x_i,\eta))^c) \le \varepsilon\frac{(C_3 + C_4\log(\det\mathbf{\Sigma}_\varepsilon))}{C_\delta\eta^2}. \tag{20}$$

Now we choose $\eta = \eta_\varepsilon := \sqrt{2\varepsilon(C_3 + C_4\log(\det\mathbf{\Sigma}_\varepsilon))/C_\delta}$ (by the rough bound (13) this η_ε tends to zero as $\varepsilon \to 0$, which permits us to apply (3)). This implies (17) with $R_1 = \frac{2C_3}{C_\delta}$ and $R_2 = \frac{2C_4}{C_\delta}$, by passing to the complement and observing that

$$\sup_{i\in\{1,\dots,n\}} \nu_\varepsilon(B(x_i,\eta_\varepsilon)) \ge \frac{1}{n}\nu_\varepsilon\left(\cup_{i\in\{1,\dots,n\}}B(x_i,\eta_\varepsilon)\right).$$

Step 3. We prove the bounds (12). As in the previous step we set

$$\eta_\varepsilon = \sqrt{\varepsilon(R_1 + R_2\log(\det\mathbf{\Sigma}_\varepsilon))}.$$

It follows from (17) that

$$\begin{aligned}\frac{1}{2n} \le \nu_\varepsilon(B(x_{i_0},\eta_\varepsilon)) &= \frac{1}{\sqrt{(2\pi\varepsilon)^d\det\mathbf{\Sigma}_\varepsilon}}\int_{B(x_{i_0},\eta_\varepsilon)}\exp\left(-\frac{1}{2\varepsilon}\left\langle x-m_\varepsilon,\mathbf{\Sigma}_\varepsilon^{-1}(x-m_\varepsilon)\right\rangle\right)dx\\ &\le \frac{1}{\sqrt{(2\pi\varepsilon)^d\det\mathbf{\Sigma}_\varepsilon}}|B(x_{i_0},\eta_\varepsilon)|\\ &\le C\frac{1}{\sqrt{\varepsilon^d\det\mathbf{\Sigma}_\varepsilon}}\eta_\varepsilon^d \le C\sqrt{\frac{(R_1+R_2\log(\det\mathbf{\Sigma}_\varepsilon))^d}{\det\mathbf{\Sigma}_\varepsilon}}.\end{aligned} \tag{21}$$

This implies that $\limsup_{\varepsilon\downarrow 0}\det\mathbf{\Sigma}_\varepsilon < C$ for some $C > 0$. In order to get a lower bound on individual eigenvalues $\mathbf{\Lambda}_\varepsilon^{(i)}$ of $\mathbf{\Sigma}_\varepsilon$, we rewrite the same integral in a slightly different way. We use the change of coordinates $y = \frac{\mathbf{P}_\varepsilon^T(x-m_\varepsilon)}{\sqrt{\varepsilon}}$, where $\mathbf{P}_\varepsilon$ is orthogonal and diagonalizes $\mathbf{\Sigma}_\varepsilon$ and observe that under this transformation $B(x^i,\eta_\varepsilon)$ is mapped into $B(\frac{x^i-m}{\sqrt{\varepsilon}},\frac{\eta_\varepsilon}{\varepsilon}) \subseteq \{y\colon |y_j - \frac{(x^i-m)}{\sqrt{\varepsilon}}| \le \frac{\eta_\varepsilon}{\sqrt{\varepsilon}}$ for $j = 1,\dots,n\}$. This yields

$$\begin{aligned}\frac{1}{2n} &\le \frac{1}{\sqrt{(2\pi)^d\det\mathbf{\Sigma}_\varepsilon}}\int_{\{|y_j-\frac{(x^i-m)}{\sqrt{\varepsilon}}|\le\frac{\eta_\varepsilon}{\sqrt{\varepsilon}}\}}\exp\left(-\frac{1}{2}\left\langle y_i,\left(\mathbf{\Lambda}_\varepsilon^{(i)}\right)^{-1}y_i\right\rangle\right)dy\\ &\le \frac{1}{\sqrt{(2\pi)^d\det\mathbf{\Sigma}_\varepsilon}}\left(\frac{2\eta_\varepsilon}{\sqrt{\varepsilon}}\right)^{d-1}\int_{\mathbb{R}}\exp\left(-\frac{|y_i|^2}{2\mathbf{\Lambda}_\varepsilon^{(i)}}\right)dy_i\\ &= \sqrt{\frac{\mathbf{\Lambda}_\varepsilon^{(i)}}{(2\pi)^d\det\mathbf{\Sigma}_\varepsilon}}\left(R_1+R_2\log(\det\mathbf{\Sigma}_\varepsilon)\right)^{\frac{d-1}{2}}\end{aligned} \tag{22}$$

for any $i \in \{1,2,\dots,d\}$. Together with uniform boundedness of $\det\mathbf{\Sigma}_\varepsilon$ this implies that $\mathbf{\Lambda}_\varepsilon^{(i)} > C'$ for some $C' > 0$. Finally,

$$\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) = \sum_{i=1}^d\mathbf{\Lambda}_\varepsilon^{(i)} = \sum_{i=1}^d\frac{\det(\mathbf{\Sigma}_\varepsilon)}{\prod_{j=1,j\ne i}^d\mathbf{\Lambda}_\varepsilon^{(j)}} \le \frac{dC}{(C')^{d-1}} < \infty. \tag{23}$$

This proves (12).

Step 4. We show that $\mathrm{dist}(m_\varepsilon, \mathscr{E}) \downarrow 0$ as $\varepsilon \downarrow 0$. On the one hand, by the upper bound on the variance in (12) and standard Gaussian concentration, we see that there exists a constant $c > 0$ such that for $\varepsilon \ll 1$ we have $\nu_\varepsilon(B(m_\varepsilon, \sqrt{\varepsilon}c)) \geq \frac{3}{4}$. On the other hand, we had already seen in (20) that for $\eta = \eta_\varepsilon$ we have

$$\nu_\varepsilon(\cup_{i=1}^n B(x_i, \eta_\varepsilon))^c) \leq \frac{1}{2},$$

and hence $B(m_\varepsilon, \sqrt{\varepsilon}c)$ must intersect at least one of the $B(x_i, \eta_\varepsilon)$. This yields for this particular index i

$$|x_i - m_\varepsilon| \leq \eta_\varepsilon + \sqrt{\varepsilon}c$$

and establishes the claim. ∎

Lemma 3.7. *Let* $\{(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)\}$ *be a sequence such that* $\limsup_{\varepsilon\downarrow 0} |m_\varepsilon| =: C_1 < \infty$ *and*

$$0 < c_2 := \liminf_{\varepsilon\downarrow 0} \lambda_{\min}(\mathbf{\Sigma}_\varepsilon) < \limsup_{\varepsilon\downarrow 0} \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) =: C_2 < \infty.$$

Then as $\varepsilon \downarrow 0$,

$$F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} + V_2(m_\varepsilon) + \frac{1}{2}\mathrm{Tr}(D^2 V_1^\varepsilon(m_\varepsilon) \cdot \mathbf{\Sigma}_\varepsilon) - \frac{1}{2}\log\left((2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon\right) - \frac{d}{2} + \log Z_{\mu,\varepsilon} + r_\varepsilon, \tag{24}$$

where $|r_\varepsilon| \leq C\varepsilon$ *with* $C = C(C_1, c_2, C_2, M_V)$ *(recall that* M_V *is the constant defined in Assumption* 2.3(A-1)*).*

Proof. The lemma follows directly from the expression (5) and the Taylor expansion. Indeed, we first expand V_2 near m_ε up to the first order and then take expectation to get

$$\mathbb{E}^{\nu_\varepsilon} V_2(x) = V_2(m_\varepsilon) + \mathbb{E}^{\nu_\varepsilon} R_\varepsilon(x)$$

with residual

$$R_\varepsilon(x) = \sum_{|\alpha|=2} \frac{(x - m_\varepsilon)^\alpha}{\alpha!} \int_0^1 \partial^\alpha V_2\left(\xi x + (1-\xi)m_\varepsilon\right)(1-\xi)^2 d\xi.$$

Thanks to the condition (A-1), one can obtain the bound

$$\begin{aligned}
\mathbb{E}^{\nu_\varepsilon} R_\varepsilon(x) &\leq \sum_{|\alpha|=2} \frac{1}{\alpha!} \max_{\xi\in[0,1]} \left\{\mathbb{E}^{\nu_\varepsilon}\left[|x - m_\varepsilon|^2 \partial^\alpha V_2\left(\xi x + (1-\xi)m_\varepsilon\right)\right]\right\} \\
&\leq \frac{M_V}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} \max_{\xi\in[0,1]} \left\{\int_{\mathbb{R}^d} |x|^2 e^{(|x|+|m_\varepsilon|)^2} \cdot e^{-\frac{1}{2\varepsilon} x^T \mathbf{\Sigma}_\varepsilon^{-1} x} dx\right\} \\
&\leq \frac{M_V}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} e^{2|m_\varepsilon|^2} \int_{\mathbb{R}^d} |x|^2 e^{-\frac{1}{2\varepsilon} x^T (\mathbf{\Sigma}_\varepsilon^{-1} - 4\varepsilon \cdot \mathbf{I}_d) x} dx \\
&= \frac{M_V \varepsilon}{\sqrt{\det \mathbf{\Sigma}_\varepsilon}} e^{2|m_\varepsilon|^2} \cdot \det(\mathbf{\Sigma}_\varepsilon^{-1} - 4\varepsilon \cdot \mathbf{I}_d)^{-1} \\
&\leq C\varepsilon,
\end{aligned} \tag{25}$$

when $\varepsilon \ll 1$. Note that in the last inequality we have used the assumption that all eigenvalues of $\mathbf{\Sigma}_\varepsilon$ are bounded from above, which ensures that for $\varepsilon \ll 1$ the matrix $\mathbf{\Sigma}_\varepsilon^{-1} - 4\varepsilon \cdot \mathbf{I}_d$ is positive definite. Hence

$$\mathbb{E}^{\nu_\varepsilon} V_2(x) = V_2(m_\varepsilon) + r_{1,\varepsilon}$$

with $r_{1,\varepsilon} \leq C\varepsilon$ as $\varepsilon \downarrow 0$. Similarly, one can take the fourth order Taylor expansion for V_1^ε near m_ε and then take expectation to obtain that

$$\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} + \frac{1}{2}\mathrm{Tr}\left(D^2 V_1^\varepsilon(m_\varepsilon) \cdot \mathbf{\Sigma}_\varepsilon\right) + r_{2,\varepsilon}$$

with $r_{2,\varepsilon} \leq C\varepsilon$. Then (24) follows directly by inserting the above equations into the expression (5). ∎

The following corollary is a direct consequence of Lemmas 3.2, 3.6, and 3.7, providing an asymptotic formula for $F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$ as $\varepsilon \downarrow 0$.

Corollary 3.8. *Let* $\{(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)\} \subset \mathbb{R}^d \times \mathcal{S}_{\geq}(\mathbb{R}, d)$ *be such that* $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. *Then for* $\varepsilon \ll 1$

$$F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} + V_2(m_\varepsilon) + \frac{1}{2}\mathrm{Tr}(D^2 V(m_\varepsilon) \cdot \mathbf{\Sigma}_\varepsilon) - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) - \frac{d}{2} + \sum_{i=1}^n \beta^i + o(1). \tag{26}$$

Remark 3.9. We do not have a bound on the convergence rate for the residual expression (26), because Lemma 3.2 does not provide a convergence rate on the $Z_{\mu,\varepsilon}$. This is because we do not impose any rate of convergence for the convergence of the x_ε^i to x^i. The bound $|r_\varepsilon| \leq C\varepsilon$ in Lemma 3.7 will be used to prove the rate of convergence for the posterior measures that arise from Bayesian inverse problems; see Theorem 5.4 and its proof.

Proof of Theorem 3.3. We first prove the liminf inequality. Let $(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$ be such that $m_\varepsilon \to m$ and $\mathbf{\Sigma}_\varepsilon \to \mathbf{\Sigma}$. We want to show that $F(m, \mathbf{\Sigma}) \leq \liminf_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$. We may assume that $\liminf_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$ since otherwise there is nothing to prove. By Lemma 3.6 this implies that $m \in \mathscr{E}$ and that $\mathbf{\Sigma}$ is positive definite. Then the liminf inequality follows from (26) and the fact that $V_1^\varepsilon \geq 0$.

Next we show the limsup inequality is true. Given $m \in \mathscr{E}, \mathbf{\Sigma} \in \mathcal{S}_>(\mathbb{R}, d)$, we want to find recovery sequences $(m_k, \mathbf{\Sigma}_k)$ such that $(m_k, \mathbf{\Sigma}_k) \to (m, \mathbf{\Sigma})$ and $\limsup_k F_{\varepsilon_k}(m_k, \mathbf{\Sigma}_k) \leq F(m, \mathbf{\Sigma})$. In fact, we set $\mathbf{\Sigma}_k = \mathbf{\Sigma}$. Moreover, by Assumption 2.3(A-4), we can choose $\{m_k\}$ to be one of the zeros of $V_1^{\varepsilon_k}$ so that $V_1^{\varepsilon_k}(m_k) = 0$ and $m_k \to m \in \mathscr{E}$. This implies that $V_2(m_k) \to V_2(m)$. Then the limsup inequality follows from (26). ∎

Proof of Corollary 3.4. First we show that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. In fact, let $\tilde{m}_\varepsilon = x_\varepsilon^1$ and $\tilde{\mathbf{\Sigma}}_\varepsilon = D^2 V_1^\varepsilon(x_\varepsilon^1)$. It follows from (26) that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(\tilde{m}_\varepsilon, \tilde{\mathbf{\Sigma}}_\varepsilon) < \infty$. According to Theorem 2.2, the convergence of minima and minimizers is a direct consequence of Lemma 3.6 and Theorem 3.3. ∎

4. Approximation by Gaussian mixtures. In the previous section we demonstrated the approximation of the target measure (1) by a Gaussian. Corollary 3.5 shows that when the measure has only one mode, this approximation is perfect in the limit $\varepsilon \to 0$: the limit Kullback–Leibler divergence tends to zero since both entropies in (11) tend to zero. However when multiple modes exist, and persist in the small ε limit, the single Gaussian is inadequate because the relative entropy term $D_{\mathrm{KL}}(\mathbf{e^i}||\boldsymbol{\beta})$ cannot be small even though the relative entropy between Gaussians tends to zero. In this section we consider the approximation of the target measure μ_ε by Gaussian mixtures in order to overcome this issue. We show that in the case of n minimizers of V_1, the approximation with a mixture of n Gaussians is again perfect as $\varepsilon \to 0$. The Gaussian mixture model is widely used in the pattern recognition and machine learning community; see the relevant discussion in [3, Chapter 9].

Let $\triangle^n$ be the standard n-simplex, i.e.,

$$\triangle^n = \left\{ \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \dots, \alpha^n) \in \mathbb{R}^n : \alpha^i \geq 0 \text{ and } \sum_{i=1}^n \alpha^i = 1 \right\}.$$

For $\xi \in (0,1)$, we define $\triangle^n_\xi = \{\boldsymbol{\alpha} = (\alpha^1, \alpha^2, \dots, \alpha^n) \in \mathbb{R}^n : \alpha^i \geq \xi\}$.

Recall that $\mathcal{A}$ is the set of Gaussian measures and define the set of Gaussian mixtures

$$\mathcal{M}_n = \left\{ \nu = \sum_{i=1}^n \alpha^i \nu^i : \nu^i \in \mathcal{A},\ \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \dots, \alpha^n) \in \triangle^n \right\}. \tag{27}$$

Also, for a fixed $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0,\infty)$ we define the set

$$\mathcal{M}_n^{\boldsymbol{\xi}} = \left\{ \nu = \sum_{i=1}^n \alpha^i \nu^i : \nu^i = N(m^i, \boldsymbol{\Sigma}^i) \in \mathcal{A} \text{ with } \min_{i \neq j} |m^i - m^j| \geq \xi_2, \right.$$
$$\left. \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \dots, \alpha^n) \in \triangle^n_{\xi_1} \right\}. \tag{28}$$

While $\mathcal{M}_n$ is the set of all convex combinations of n Gaussians taken from $\mathcal{A}$; the set $\mathcal{M}_n^{\boldsymbol{\xi}}$ can be seen as an "effective" version of $\mathcal{M}_n$, in which each Gaussian component plays an active role, and no two Gaussians share a common center.

Consider the problem of minimizing $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$ within $\mathcal{M}_n$ or $\mathcal{M}_n^{\boldsymbol{\xi}}$. Since the sets $\mathcal{M}_n$ and $\mathcal{M}_n^{\boldsymbol{\xi}}$ are both closed with respect to weak convergence, we have the following existence result, whose proof is similar to Theorem 3.1 and is omitted.

Theorem 4.1. *Consider the measure μ_ε given by (1) with fixed $\varepsilon > 0$ and the problem of minimizing the functional*

$$\nu \mapsto D_{KL}(\nu||\mu_\varepsilon) \tag{29}$$

from the set $\mathcal{M}_n$ or from the set $\mathcal{M}_n^{\boldsymbol{\xi}}$ with some fixed $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0,\infty)$. In both cases, there exists at least one minimizer to the functional (29).

Now we continue to investigate the asymptotic behavior of the Kullback–Leibler approximations based on Gaussian mixtures. To that end, we again parametrize a measure ν in the set $\mathcal{M}_n$ or $\mathcal{M}_n^{\boldsymbol{\xi}}$ by the weights $\boldsymbol{\alpha} = (\alpha^1, \alpha^2, \dots, \alpha^n)$ as well as the n means as well

as the n covariances matrices. Similar to the previous section we need to chose the right scaling in our Gaussian mixtures to reflect the typical size of fluctuations of μ_ε. Thus for $\mathbf{m} = (m^1, m^2, \dots, m^n)$ and $\mathbf{\Sigma} = (\mathbf{\Sigma}^1, \mathbf{\Sigma}^2, \dots, \mathbf{\Sigma}^n)$. we set

$$\nu_\varepsilon = \sum_{i=1}^{n} \alpha^i N(m^i, \varepsilon \mathbf{\Sigma}^i). \tag{30}$$

We can view $D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon)$ as a functional of $(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{\Sigma})$ and study the Γ-convergence of the resulting functional. For that purpose, we need to restrict our attention to finding the best Gaussian mixtures within $\mathcal{M}_n^{\boldsymbol{\xi}}$ for some $\boldsymbol{\xi} \in (0,1) \times (0, \infty)$. The reasons are the following. First, we require individual Gaussian measures ν^i to be active (i.e., $\alpha^i > \xi_1 > 0$) because $D_{\mathrm{KL}}(\nu_\varepsilon, \mu_\varepsilon)$, as a family of functionals of $(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{\Sigma})$ indexed by ε, is not compact if we allow some of the α^i to vanish. In fact, if $\alpha_\varepsilon^i = 0$ for some $i \in 1, 2, \dots, n$, then $D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon)$ is independent of m_ε^i and $\mathbf{\Sigma}_\varepsilon^i$. In particular, if $|m_\varepsilon^i| \wedge |\mathbf{\Sigma}_\varepsilon^i| \to \infty$ while $|m_\varepsilon^j| \vee |\mathbf{\Sigma}_\varepsilon^j| < \infty$ for all the j's such that $j \neq i$, then it still holds that $\limsup_{\varepsilon \downarrow 0} D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon) < \infty$. Second, it makes more sense to assume that the individual Gaussian means stay apart from each other (i.e., $\min_{i \neq j} |m^i - m^j| \geq \xi_2 > 0$) since we primarily want to locate different modes of the target measure. Moreover, it seems impossible to identify a sensible Γ-limit without such an assumption; see Remark 4.7.

Recall that the measure ν has the form (30). Let $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0, \infty)$ be fixed. In view of these considerations it is useful to define

$$S_{\boldsymbol{\xi}} = \{(\boldsymbol{\alpha}, \mathbf{m}) \in \triangle_{\xi_1}^n \times \mathbb{R}^{nd} : \min_{i \neq j} |m^i - m^j| \geq \xi_2\}.$$

We define the functional

$$G_\varepsilon(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{\Sigma}) := \begin{cases} D_{\mathrm{KL}}(\nu || \mu_\varepsilon) & \text{if } (\boldsymbol{\alpha}, \mathbf{m}) \in S_{\boldsymbol{\xi}}, \\ +\infty & \text{otherwise.} \end{cases} \tag{31}$$

By the definition of the Kullback–Leibler divergence, if $(\boldsymbol{\alpha}, \mathbf{m}) \in S_{\boldsymbol{\xi}}$, then

$$G_\varepsilon(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{\Sigma}) = \int \rho(x) \log \rho(x) dx + \frac{1}{\varepsilon} \mathbb{E}^\nu V_1^\varepsilon(x) + \mathbb{E}^\nu V_2(x) + \log Z_{\mu,\varepsilon}, \tag{32}$$

where ρ is the probability density function of ν.

Recall the Γ-limit F defined in (9). Then we have the following Γ-convergence result.

Theorem 4.2. *The Γ-limit of G_ε is*

$$\begin{aligned} G(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{\Sigma}) := \sum_{i=1}^{n} \alpha^i D_{KL}\left(N(m^i, \mathbf{\Sigma}^i) \;||\; N(m^i, (D^2 V_1(m^i))^{-1})\right) \\ + D_{KL}(\boldsymbol{\alpha} \;||\; \boldsymbol{\beta}) \end{aligned} \tag{33}$$

if $(\boldsymbol{\alpha}, \mathbf{m}) \in S_{\boldsymbol{\xi}}$ and $m^i \in \mathscr{E}$ and ∞ otherwise.

Remark 4.3. The right-hand side of G consists of two parts: the first part is a weighted relative entropy which measures the discrepancy between two Gaussians, and the second part

is the relative entropy between sums of Dirac masses at $\{x^j\}_{j=1}^n$ with weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. This has the same spirit as the entropy splitting used in [21, Lemma 2.4].

Before we prove Theorem 4.2, we consider the minimization of the limit functional G. First let ξ_1, ξ_2 be such that

$$\xi_1 > 0, \quad 0 < \xi_2 \le \min_{i \neq j} |x^i - x^j|, \tag{34}$$

where $\{x^i\}_{i=1}^n$ are the minimizers of V_1. To minimize G, without loss of generality, we may choose $m^i = \overline{m}^i := x^i$. Then the weighted relative entropy in the first term in the definition (33) of G vanishes if we set $\boldsymbol{\Sigma}^i = \overline{\boldsymbol{\Sigma}}^i := D^2 V_1(x^i)^{-1}$. The relative entropy of the weights also vanishes if we choose the weight $\boldsymbol{\alpha} = \overline{\boldsymbol{\alpha}} := \boldsymbol{\beta}$. To summarize, the minimizer $(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}})$ of G is given by

$$\overline{m}^i = x^i, \ \overline{\boldsymbol{\Sigma}}^i = D^2 V_1(x^i)^{-1}, \overline{\alpha}^i = \beta^i, \tag{35}$$

and $G(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}}) = 0$. The following corollary is a direct consequence of the Γ-convergence of G_ε.

Corollary 4.4. *Let $\{(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)\}$ be a family of minimizers of $\{G_\varepsilon\}$. Then there exists a subsequence $\{\varepsilon_k\}$ such that $(\boldsymbol{\alpha}_{\varepsilon_k}, \mathbf{m}_{\varepsilon_k}, \boldsymbol{\Sigma}_{\varepsilon_k}) \to (\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}})$ and that $G_{\varepsilon_k}(\boldsymbol{\alpha}_{\varepsilon_k}, \mathbf{m}_{\varepsilon_k}, \boldsymbol{\Sigma}_{\varepsilon_k}) \to G(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}})$. Moreover, $(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}})$ is a minimizer of G and $G(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{m}}, \overline{\boldsymbol{\Sigma}}) = 0$.*

For a non-Gaussian measure μ_ε with multiple modes, i.e., $n > 1$ in the Assumption 2.3, we have seen in Corollary 3.5 that the Kullback–Leibler divergence between μ_ε and the best Gaussian measure selected from $\mathcal{A}$ remains positive as $\varepsilon \downarrow 0$. However, this gap is filled by using Gaussian mixtures, namely, with ν_ε being chosen as the best Gaussian mixture, the Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$.

Similarly to the proof of Theorem 3.3, Theorem 4.2 follows directly from Corollary 4.8 below, whose proof requires several lemmas. We start by showing the compactness of $\{G_\varepsilon\}$.

Lemma 4.5. *Let G_ε be defined by (31). Fix $\boldsymbol{\xi} = (\xi_1, \xi_2)$ satisfying the condition (34). Let $\{(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)\}$ be a sequence in $S_{\boldsymbol{\xi}} \times \mathcal{S}_{\ge}(\mathbb{R}, d)$ such that*

$$\limsup_{\varepsilon \downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon) < \infty.$$

Then

$$\liminf_{\varepsilon \downarrow 0} \min_i \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon^i) > 0, \ \limsup_{\varepsilon \downarrow 0} \max_i \mathrm{Tr}(\boldsymbol{\Sigma}_\varepsilon^i) < \infty \tag{36}$$

and $\mathrm{dist}(m_\varepsilon^i, \mathscr{E}) \downarrow 0$ as $\varepsilon \downarrow 0$. In particular, for any i, there exists $j = j(i) \in \{1, 2, \dots, n\}$ and a subsequence $\{m_k^i\}_{k \in \mathbb{N}}$ of $\{m_\varepsilon^i\}$ such that $m_k \to x_j$ as $k \to \infty$.

Proof. We write $M = \limsup_{\varepsilon \downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$ and

$$\nu_\varepsilon = \sum_{i=1}^n \alpha_\varepsilon^i \nu_\varepsilon^i,$$

where $\nu_\varepsilon^i = N(m_\varepsilon^i, \varepsilon\mathbf{\Sigma}_\varepsilon^i)$. Then we get

$$\begin{aligned} D_{KL}(\nu_\varepsilon||\mu_\varepsilon) &= \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log\left(\sum_i \alpha_\varepsilon^i \frac{d\nu_\varepsilon^i}{d\mu_\varepsilon}\right) \\ &\geq \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log\left(\alpha_\varepsilon^j \, \frac{d\nu_\varepsilon^j}{d\mu_\varepsilon}\right) \\ &= \sum_{j=1}^n \alpha_\varepsilon^j \log(\alpha_\varepsilon^j) + \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log\left(\frac{d\nu_\varepsilon^j}{d\mu_\varepsilon}\right) \\ &= \sum_{j=1}^n \alpha_\varepsilon^j \log(\alpha_\varepsilon^j) + \sum_{j=1}^n \alpha_\varepsilon^j D_{KL}(\nu_\varepsilon^j||\mu_\varepsilon), \end{aligned}$$

where the inequality follows simply from the monotonicity of the logarithmic function. As each of term $D_{KL}(\nu_\varepsilon^j||\mu_\varepsilon)$ is nonnegative, this implies the bound

$$D_{KL}(\nu_\varepsilon^j||\mu_\varepsilon) \leq \frac{1}{\alpha_\varepsilon^j}\left(M - n \min_{\alpha\in[0,1]} \alpha\log\alpha\right).$$

Using the lower bound $\alpha_\varepsilon^j > \xi_1$ which holds by assumption we get a uniform upper bound on $D_{KL}(\nu_\varepsilon^j||\mu_\varepsilon)$ which in turn permits to invoke Lemma 3.6. ∎

Lemma 4.6. *Let $\{(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon)\}$ be a sequence in $S_{\boldsymbol{\xi}} \times S_{\geq}(\mathbb{R}, d)$ with $\boldsymbol{\xi} = (\xi_1, \xi_2)$ satisfying (34) and such that*

$$c_1 \leq \liminf_{\varepsilon\downarrow 0} \min_i \lambda_{\min}(\mathbf{\Sigma}_\varepsilon^i) < \limsup_{\varepsilon\downarrow 0} \max_i |m_\varepsilon^i| \vee \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon^i) \leq C_1 < \infty.$$

Then

$$\begin{aligned} (37)\quad G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) = &\sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + V_2(m_\varepsilon^i) + \frac{1}{2}\mathrm{Tr}(D^2 V_1^\varepsilon(m_\varepsilon^i)\cdot\mathbf{\Sigma}_\varepsilon^i) - \frac{1}{2}\log\left(\det\mathbf{\Sigma}_\varepsilon^i\right)\right) \\ &+ \sum_{i=1}^n \alpha_\varepsilon^i \log\alpha_\varepsilon^i - \frac{d}{2} + \log Z_{\mu,\varepsilon} + r_\varepsilon, \end{aligned}$$

where $r_\varepsilon \leq C\varepsilon$ with $C = C(c_1, C_1, M_V, \xi_2)$.

Proof. By assumption, we know from (32) that

$$G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \int \rho_\varepsilon(x)\log\rho_\varepsilon(x)dx + \frac{1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) + \log Z_{\mu,\varepsilon},$$

where $\rho_\varepsilon = \sum_{i=1}^n \alpha_\varepsilon^i \rho_\varepsilon^i$ is the probability density of the measure ν_ε. First of all, applying the same Taylor expansion arguments used to obtain (24), one can deduce that

$$(38)\quad \frac{1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) = \sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + \frac{1}{2}\mathrm{Tr}\left(\nabla^2 V_1^\varepsilon(m_\varepsilon^i)\cdot\mathbf{\Sigma}_\varepsilon^i\right) + V_2(m_\varepsilon^i)\right) + r_{1,\varepsilon}$$

with $r_{1,\varepsilon} \leq C\varepsilon$ and $C = C(C_1, c_1, M_V)$. Next, we claim that the entropy of ρ_ε can be rewritten as

$$\int \rho_\varepsilon(x) \log \rho_\varepsilon(x) dx = \sum_{i=1}^n \alpha_\varepsilon^i \left(\int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i \right) + r_{2,\varepsilon}, \tag{39}$$

where $r_{2,\varepsilon} \leq e^{-\frac{C}{\varepsilon}}$ when $\varepsilon \ll 1$ with the constant $C = C(C_1, c_2, \xi_2)$. By definition,

$$\int \rho_\varepsilon(x) \log \rho_\varepsilon(x) dx = \sum_{i=1}^n \alpha_\varepsilon^i \int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx,$$

so it suffices to show that for each $i \in \{1, \dots, n\}$ we have

$$\int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx = \int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i + r_{2,\varepsilon} \tag{40}$$

with $r_{2,\varepsilon} \leq e^{-\frac{C}{\varepsilon}}$. Indeed, the monotonicity of the logarithmic function yields

$$\int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx \geq \int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i. \tag{41}$$

In order to show the matching lower bound we first recall that the means m_ε^i of the ν_ε^i are well separated by assumption, $\min_{j \neq i} |m_\varepsilon^i - m_\varepsilon^j| > \xi_2$. Let $\delta \ll \frac{\xi}{2}$ be fixed below and set $B_\delta^i = B(m_\varepsilon^i, \delta)$. Then we write

$$\begin{aligned} \int \rho_\varepsilon^i \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \right) &= \int \rho_\varepsilon^i \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) + \int_{B_\delta^i} \rho_\varepsilon^i \left(\log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \right) - \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) \right) \\ &\quad + \int_{(B_\delta^i)^c} \rho_\varepsilon^i \left(\log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \right) - \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) \right) \\ &=: \left(\int \rho_\varepsilon^i \log \rho_\varepsilon^i + \log \alpha_\varepsilon^i \right) + E_\varepsilon^1 + E_\varepsilon^2. \end{aligned} \tag{42}$$

We first show that the error term E_ε^2 is exponentially small. To that end, we first drop the exponential term in the Gaussian density to obtain the crude bound

$$\log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \right) \leq \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \frac{1}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon^j}} \right) \leq \frac{d}{2} \log \varepsilon^{-1} + C, \tag{43}$$

where in the second inequality we use the fact that $\det \mathbf{\Sigma}_\varepsilon^i$ is bounded away from zero, which has been established in (36). Moreover, by definition we have

$$-\log\left(\alpha_\varepsilon^i \rho_\varepsilon^i\right) \le \frac{d}{2}\log\varepsilon^{-1} + C + \frac{|x-m_\varepsilon^i|^2}{\varepsilon}. \tag{44}$$

Plugging bounds (43) and (43) in and using Gaussian concentration as well as the lower bound on $\lambda_{\min}$ established in Lemma 4.5,

$$E_\varepsilon^2 \le \int_{(B_\delta^i)^c} \rho_\varepsilon^i(x)\left(\frac{d}{2}\log\varepsilon^{-1} + C + \frac{|x-m_\varepsilon^i|^2}{\varepsilon}\right)dx \le C\big(\log\varepsilon^{-1} + \varepsilon^{-1}\big)e^{-\frac{C\delta}{\varepsilon}} \tag{45}$$

when $\varepsilon \ll 1$. Next, we want to bound E_ε^1. Notice that $m_\varepsilon^j \to m^j$ for $j = 1, \dots, n$; hence if $x \in B_\delta^i$ and if $\delta < \xi_1$, then $|x - m_\varepsilon^j| > \xi_1 - \delta$ for any $j \ne i$ when $\varepsilon \ll 1$. As a consequence,

$$\int_{B_\delta^i} \sum_{j=1, j\ne i}^n \alpha_\varepsilon^j \rho_\varepsilon^j \le C\varepsilon^{-\frac{d}{2}} e^{-\frac{C(\xi_1-\delta)^2}{\varepsilon}}. \tag{46}$$

This together with the elementary inequality

$$\log(x+y) = \log(x) + \int_x^{x+y} \frac{1}{t}dt \le \log x + \frac{y}{x}$$

for $x, y > 0$ implies

$$\begin{aligned} E_\varepsilon^1 &= \int_{B_\delta^i} \rho_\varepsilon^i\Big(\log\Big(\alpha_\varepsilon^i\rho_\varepsilon^i + \sum_{j=1,j\ne i}^n \alpha_\varepsilon^j\rho_\varepsilon^j\Big) - \log\left(\alpha_\varepsilon^i\rho_\varepsilon^i\right)\Big) \\ &\le \int_{B_\delta^i} \frac{\sum_{j=1,j\ne i}^n \alpha_\varepsilon^j\rho_\varepsilon^j}{\alpha_\varepsilon^i} \\ &\le C\delta^d\varepsilon^{-\frac{d}{2}} e^{-\frac{C(\xi_1-\delta)^2}{\varepsilon}}, \end{aligned} \tag{47}$$

where we used that α_ε^i is bounded below from zero. Hence (40) follows directly from (41)–(47).

Finally, (37) follows from combining (38), (39) and the identity

$$\int \rho_\varepsilon^i(x)\log\rho_\varepsilon^i(x)dx = -\frac{1}{2}\log\left((2\pi\varepsilon)^d \det\mathbf{\Sigma}_\varepsilon^i\right) - \frac{d}{2}. \qquad \blacksquare$$

Remark 4.7. The assumption that $\min_{j\ne i}|m_\varepsilon^i - m_\varepsilon^j| > \xi_2 > 0$ is the crucial condition that allows us to express the entropy of the Gaussian mixture in terms of the mixture of entropies of individual Gaussian (i.e., (39)), leading to the asymptotic formula (37). Neither formula (39) nor (24) is likely to be true without such an assumption since the cross entropy terms are not negligible.

The following corollary immediately follows from Lemma 4.6 by plugging in the Laplace approximation of the normalization constant $Z_{\mu,\varepsilon}$ given in Lemma 3.2 and rearranging the terms.

Corollary 4.8. *Suppose that* $\{(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)\}$ *satisfy the same assmption as in Lemma* 4.5. *If* $\limsup_{\varepsilon\downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon) < \infty$, *then*

$$G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon) = \sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + V_2(m_\varepsilon^i) - \frac{d}{2} + \frac{1}{2}\mathrm{Tr}(D^2 V_1^\varepsilon(m_\varepsilon^i) \cdot \boldsymbol{\Sigma}_\varepsilon^i) \right) + \sum_{i=1}^n \alpha_\varepsilon^i \left(\log \alpha_\varepsilon^i - \frac{1}{2}\log\left(\det \boldsymbol{\Sigma}_\varepsilon^i\right) + \log\left(\sum_{j=1}^n \beta^j \right) \right) + o(1). \tag{48}$$

Remark 4.9. Similarly to the discussion in Remark 3.9, the residual in (48) is here demonstrated to be of order $o(1)$, but the quantitative bound that $|r_\varepsilon| \le C\varepsilon$ in (37) can be used to extract a rate of convergence. This can be used to study the limiting behavior of posterior measures arising from Bayesian inverse problems when multiple modes are present; see the next section.

5. Applications in Bayesian inverse problems. Consider the inverse problem of recovering $x \in \mathbb{R}^d$ from the noisy data $y \in \mathbb{R}^d$, where y and x are linked through the equation

$$y = G(x) + \eta. \tag{49}$$

Here G is called the forward operator which maps from $\mathbb{R}^d$ into itself, and $\eta \in \mathbb{R}^d$ represents the observational noise. We take a Bayesian approach to solving the inverse problem. The main idea is to first model our knowledge about x with a prior probability distribution, leading to a joint distribution on (x, y) once the probabilistic structure on η is defined. We then update the prior based on the observed data y; specifically we obtain the posterior distribution μ^y which is the conditional distribution of x given y and is the solution to the Bayesian inverse problem. From this posterior measure one can extract information about the unknown quantity of interest. We remark that since G is nonlinear in general, the posterior is generally not Gaussian even when the noise and prior are both assumed to be Gaussian. A systematic treatment of the Bayesian approach to inverse problems may be found in [27].

In Bayesian statistics there is considerable interest in the study of the asymptotic performance of posterior measures from a frequentist perspective; this is often formalized as the *posterior consistency*. To define this precisely, consider a sequence of observations $\{y_j\}_{j\in\mathbb{N}}$, generated from the truth $x^\dagger$ via

$$y_j = G(x^\dagger) + \eta_j, \tag{50}$$

where $\{\eta_j\}_{j\in\mathbb{N}}$ is a sequence of random noises. This may model a statistical experiment with increasing amounts of data or with vanishing noise. In either case, posterior consistency refers to concentration of the posterior distribution around the truth as the data quality increases. For parametric statistical models, Doob's consistency theorem [28, Theorem 10.10] guarantees posterior consistency under the identifiability assumption about the forward model. For nonparametric models, in which the parameters of interest lie in infinite dimensional spaces, the corresponding posterior consistency is a much more challenging problem. Schwartz's theorem [25, 2] provides one of the main theoretical tools to prove posterior consistency in infinite

dimensional space, which replaces identifiability by a stronger assumption on testability. The posterior contraction rate, quantifying the speed that the posterior contracts to the truth, has been determined in various Bayesian statistical models (see [10, 26, 7]). In the context of the Bayesian inverse problem, the posterior consistency problem has mostly been studied to date for linear inverse problems with Gaussian priors [16, 1]. The recent paper [29] studied posterior consistency for a specific nonlinear Bayesian inverse problem, using the stability estimate of the underlying inverse problem together with posterior consistency results for the Bayesian regression problem.

In this section, our main interest is not in the consistency of posterior distribution but in characterizing in detail its asymptotic behavior. We will consider two limit processes in (50): the small noise limit and the large data limit. In the former case, we assume that the noise $\eta_i = \frac{1}{\sqrt{i}}\eta$, where η is distributed according to the standard normal $N(0, \mathbf{I}_d)$, and we consider the data $\mathbf{y}_N$ given by the most accurate observation, i.e., $\mathbf{y}_N = y_N$. In the later case, the sequence $\{\eta_i\}_{i\in\mathbb{N}}$ is assumed to be independent identically distributed (i.i.d) according to the standard normal and we accumulate the observations so that the data $\mathbf{y}_N = \{y_1, y_2, \dots, y_N\}$. In addition, assume that the prior distribution is μ_0 which has the density

$$\mu_0(dx) = \frac{1}{Z_0} e^{-V_0(x)} dx$$

with the normalization constant $Z_0 > 0$. Since the data and the posterior are fully determined by the noise $\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \eta$ or $\boldsymbol{\eta} = \{\eta_i\}_{i\in\mathbb{N}}$, we denote the posterior by $\mu_N^{\boldsymbol{\eta}}$ to indicate the dependence. By using Bayes's formula, we calculate the posterior distribution for both limiting cases below.

- Small noise limit:

$$\begin{aligned} \mu_N^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_{N,1}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}\left|y_n - G(x)\right|^2\right)\mu_0(dx) \\ &= \frac{1}{Z_{N,1}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}\left|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta\right|^2\right)\mu_0(dx). \end{aligned} \tag{51}$$

- Large data limit:

$$\begin{aligned} \mu_N^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left|y_i - G(x)\right|^2\right)\mu_0(dx) \\ &= \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left|G(x^\dagger) - G(x) + \eta_i\right|^2\right)\mu_0(dx). \end{aligned} \tag{52}$$

In both cases, we are interested in the limiting behavior of the posterior distribution $\mu_N^{\boldsymbol{\eta}}$ as $N \to \infty$. For doing so, we assume the forward operator G satisfies one of the following assumptions.

Assumption 5.1.

(i) *$G \in C^4(\mathbb{R}^d; \mathbb{R}^d)$ and $G(x) = G(x^\dagger)$ implies $x = x^\dagger$. Moreover, G is a C^1-diffeomorphism in the neighborhood of $x^\dagger$.*

(ii) $G \in C^4(\mathbb{R}^d;\mathbb{R}^d)$ *and the zero set of the equation* $G(x) = G(x^\dagger)$ *is* $\{x_i^\dagger\}_{i=1}^n$. *Moreover* $x_1^\dagger = x^\dagger$ *and* G *is a* C^1*-diffeomorphism in the neighborhood of* $x_i^\dagger$.

The following model problem gives a concrete example where these assumptions are satsified.

Model problem. Consider the following one-dimensional elliptic problem:

$$-u''(x) + \exp(q(x))u(x) = f(x), \qquad x \in (0,1), \tag{53}$$
$$u(0) = u(1) = 0.$$

Here we assume that $q, f \in L^\infty(0,1)$ and that f is positive on $(0,1)$. The inverse problem of interest is to find q from the knowledge of the solution u. We restrict ourselves to a finite dimensional version of (53), which comes from the finite difference discretization

$$-\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + e^{q_k}u_k = f_k, \qquad k = 1, 2, \dots, M, \tag{54}$$
$$u_0 = u_{M+1} = 0.$$

Here u_k, f_k, and q_k are approximations to $u(x_k)$, $f(x_k)$, and $q(x_k))$ with $x_k = k/M, k = 1, \dots, M$, and $h = 1/(M+1)$. The corresponding finite dimensional inverse problem becomes finding the vector $\mathbf{q} = \{q_k\}_{k=1}^M$ from the solution vector $\mathbf{u} = \{u_k\}_{k=1}^M$ given the right side $\mathbf{f} = \{f_k\}_{k=1}^M$. For ease of notation, let us denote by $\mathcal{A}$ the matrix representation of the one-dimensional discrete Laplacian, i.e., $\mathcal{A}_{ii} = 2/h^2$ for $i = 1, 2, \dots, M$ and $\mathcal{A}_{ij} = -1/h^2$ when $|i-j| = 1$. Let $\mathcal{Q}$ be the diagonal matrix with $\mathcal{Q}_{ii} = e^{q_i}, i = 1, 2, \dots, M$. With these notations, we can write the forward map G as

$$G : \mathbf{q} \in \mathbb{R}^M \to \mathbf{u} \in \mathbb{R}^M, \qquad \mathbf{u} = G(\mathbf{q}) = (\mathcal{A} + \mathcal{Q})^{-1}\mathbf{f}.$$

Note that both $\mathcal{A}$ and $\mathcal{Q}$ are positive definite so that $(\mathcal{A} + \mathcal{Q})$ is invertible. We now discuss this forward map, and variants on it, in relation to Assumption 5.1.

First consider Assumption 5.1(i). First, G is smooth in $\mathbf{q}$ since $\mathcal{Q}$ depends smoothly on $\mathbf{q}$. In particular, for any fixed $\mathbf{q} \in \mathbb{R}^M$ with corresponding solution vector $\mathbf{u}$, a direct calculation shows that the derivative matrix $D_\mathbf{q}G$ of the forward map G is given by

$$D_\mathbf{q}G = (\mathcal{A} + \mathcal{Q})^{-1}\mathcal{U}\mathcal{Q}.$$

Here $\mathcal{U}$ is a diagonal matrix with the diagonal vector $\mathbf{u}$. Due to our assumption that f_k are positive, it follows from the (discrete) maximum principle that the u_k are also positive, which in turn implies that $\mathcal{U}$ is invertible. Consequently, the matrix $D_\mathbf{q}G$ is invertible and

$$D_\mathbf{q}G^{-1} = \mathcal{Q}^{-1}\mathcal{U}^{-1}(\mathcal{A} + \mathcal{Q}).$$

According to the inverse function theorem, the map $G : \mathbb{R}^M \to \mathbb{R}^M$ is invertible at every $\mathbf{q} \in \mathbb{R}^M$ and its inverse $G^{-1}(\mathbf{u})$ is smooth in $\mathbf{u}$. Therefore Assumption 5.1(i) is fulfilled for any $x^\dagger = \mathbf{q}^\dagger \in \mathbb{R}^M$. The problem (53) can be modified slightly so that Assumption 5.1(ii) is satisfied. In fact, consider the problem (53) with the coefficient $\exp(q)$ replaced by q^2. Then

Assumption 5.1(ii) is satisfied for any $x^\dagger = \mathbf{q}^\dagger$ without zero entries. More specifically, the resulting forward map in this case is still smooth, but the equation $G(\mathbf{q}) = G(\mathbf{q}^\dagger)$ has $n = 2^M$ solutions $\{\mathbf{q}_i^\dagger\}_{i=1}^{2^M}$ corresponding to the fact that $\mathbf{q}$ is only determined up to a sign in each entry. Moreover, if $\mathbf{q}^\dagger$ has no zero entry, G^{-1} is smooth near each of $\mathbf{q}_i^\dagger$.

We divide our exposition below according to whether the noise is fixed or is considered as a random variable. For a fixed realization of noise $\boldsymbol{\eta} = \eta$, by applying the theory developed in the previous section, we show the asymptotic normality for $\mu_N^{\boldsymbol{\eta}}$ in the small noise limit. Furthermore, we obtain a BvM type theorem for $\mu_N^{\boldsymbol{\eta}}$ with respect to both limit processes, small noise and large data.

5.1. Asymptotic normality. In this subsection, we assume that the data is generated from the truth $x^\dagger$ and a single realization of the Gaussian noise $\eta^\dagger$, i.e.,

$$y = G(x^\dagger) + \frac{1}{\sqrt{N}}\eta^\dagger.$$

Then the resulting posterior distribution $\mu_N^{\boldsymbol{\eta}}$ has a density of the form

$$\begin{aligned}\mu_N^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_N^{\boldsymbol{\eta}}}\exp\left(-\frac{N}{2}|y - G(x)|^2 - V_0(x)\right)dx \\ &= \frac{1}{Z_N^{\boldsymbol{\eta}}}\exp\left(-\frac{N}{2}|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta^\dagger|^2 - V_0(x)\right)dx,\end{aligned} \tag{55}$$

where $Z_N^{\boldsymbol{\eta}}$ is the normalization constant. Notice that $\mu_N^{\boldsymbol{\eta}}$ has the same form as the measure defined in (1) with $\varepsilon = \frac{1}{N}, V_1^\varepsilon(x) = V_1^N(x) := \frac{1}{2}|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta^\dagger|^2$, and $V_2(x) = V_0(x)$.

Suppose that $V_0 \in C^2(\mathbb{R}^d;\mathbb{R})$ and that G satisfies one of the assumptions in Assumption (5.1). Then the potentials V_1^ε and V_2 satisfy Assumption 2.3. In particular, we have $V_1^\varepsilon(x) \to V_1(x) := \frac{1}{2}|G(x^\dagger) - G(x)|^2$ for any $x \in \mathbb{R}^d$ and that $D^2V_1(x_i^\dagger) = DG(x_i^\dagger)^T DG(x_i^\dagger)$. Recall the set of Gaussian measures $\mathcal{A}$ and the set of Gaussian mixtures $\mathcal{M}_n$ and $\mathcal{M}_n^{\boldsymbol{\xi}}$ (defined in (27) and (28)). Again, we set $\boldsymbol{\xi} = (\xi_1, \xi_2)$ such that $\xi_1 \in (0,1)$ and $\min_{i\neq j}|x^i - x^j| \geq \xi_2 > 0$. The following theorem concerning the asymptotic normality of $\mu_N^{\boldsymbol{\eta}}$ is a direct consequence of Corollaries 3.4 and 4.4.

Theorem 5.2.

(i) *Let $V_0 \in C^2(\mathbb{R}^d;\mathbb{R})$ and G satisfy Assumption 5.1(i). Given any $N \in \mathbb{N}$, let $\nu_N = N(m_N, \frac{1}{N}\boldsymbol{\Sigma}_N) \in \mathcal{A}$ be a minimizer of the functional $\nu \mapsto D_{KL}(\nu||\mu_N^{\boldsymbol{\eta}})$ within $\mathcal{A}$. Then $D_{KL}(\nu_N||\mu_N^{\boldsymbol{\eta}}) \downarrow 0$ as $N \to \infty$. Moreover, $m_N \to x^\dagger$ and $\boldsymbol{\Sigma}_N \to \left(DG(x^\dagger)^T DG(x^\dagger)\right)^{-1}$.*

(ii) *Let $V_0 \in C^2(\mathbb{R}^d;\mathbb{R})$ and G satisfy Assumption 5.1(ii). Given any $N \in \mathbb{N}$, let $\nu_N \in \mathcal{M}_n^{\boldsymbol{\xi}}$ be a minimizer of the functional $\nu \mapsto D_{KL}(\nu||\mu_N^{\boldsymbol{\eta}})$ within $\mathcal{M}_n^{\boldsymbol{\xi}}$. Let $\nu_N = \sum_{i=1}^n \alpha_N^i \nu_N^i$ with $\nu_N^i = N(m_N^i, \frac{1}{N}\boldsymbol{\Sigma}_N^i)$. Then it holds that as $N \to \infty$*

$$m_N^i \to x_i^\dagger, \boldsymbol{\Sigma}_N^i \to \left(DG(x_i^\dagger)^T DG(x_i^\dagger)\right)^{-1} \text{ and } \alpha_N^i \to \frac{\left[\det DG(x_i^\dagger)\right]^{-1} \cdot e^{-V_0(x_i^\dagger)}}{\sum_{j=1}^n \left[\det DG(x_j^\dagger)\right]^{-1} \cdot e^{-V_0(x_j^\dagger)}}.$$

Theorem 5.2(i) states that the measure $\mu_N^{\boldsymbol{\eta}}$ is asymptotically Gaussian when certain uniqueness and stability properties hold in the inverse problem. Moreover, in this case, the asymptotic Gaussian distribution is fully determined by the truth and the forward map and is independent of the prior. In the case where the uniqueness fails, but the data only corresponds to a finite number of unknowns, Theorem 5.2(ii) demonstrates that the measure $\mu_N^{\boldsymbol{\eta}}$ is asymptotically a Gaussian mixture, with each Gaussian mode independent of the prior. However, prior beliefs affect the proportions of the individual Gaussian components within the mixture; more precisely, the unnormalized weights of each Gaussian mode are proportional to the values of the prior evaluated at the corresponding unknowns.

Remark 5.3. In general, when $\{\eta_i\}_{i\in\mathbb{N}}$ is a sequence of fixed realizations of the normal distribution, Theorem 5.2 does not hold for the measure $\mu_N^{\boldsymbol{\eta}}$ defined in (52) in the large data case. However, we will show that $D_{\mathrm{KL}}(\nu_N||\mu_N^{\boldsymbol{\eta}})$ will converge to zero in some average sense; see Theorem 5.4.

5.2. A BvM type result. The asymptotic Gaussian phenomenon in Theorem 5.2 is very much in the same spirit as the celebrated BvM theorem [28]. This theorem asserts that for a certain class of regular priors, the posterior distribution converges to a Gaussian distribution, independently of the prior, as the sample size tends to infinity. Let us state the BvM theorem more precisely in the i.i.d. case. Consider observing a set of i.i.d. samples $\mathbf{X}^N := \{X^1, X^2, \dots, X^N\}$, where X^i is drawn from distribution P_θ, indexed by an unknown parameter $\theta \in \Theta$. Let P_θ^N be the law of $\mathbf{X}^N$. Let Π be the prior distribution on θ and denote by $\Pi(\cdot|\mathbf{X}^N)$ the resulting posterior distribution. The BvM theorem is concerned with the behavior of the posterior $\Pi(\cdot|\mathbf{X}^N)$ under the frequentist assumption that X^i is drawn from some true model P_{θ_0}. A standard finite dimensional BvM result (see, e.g., [28, Theorem 10.1]) states that, under certain conditions on the prior Π and the model P_θ, as $N \to \infty$

$$d_{TV}\left(\Pi(\theta|\mathbf{X}^N), N\left(\hat{\theta}_N, \frac{1}{N} I_{\theta_0}^{-1}\right)\right) \xrightarrow{P_{\theta_0}^N} 0, \tag{56}$$

where $\hat{\theta}_N$ is an efficient estimator for θ, I_θ is the Fisher information matrix of P_θ, and d_{TV} represents the total variation distance. As an important consequence of the BvM result, Bayesian credible sets are asymptotically equivalent to frequentist confidence intervals. Moreover, it has been proved that the optimal rate of convergence in the BvM theorem is $O(1/\sqrt{N})$; see, for instance, [6, 13]. This means that for any $\delta > 0$, there exists $M = M(\delta) > 0$ such that

$$P_{\theta_0}^N\left(\mathbf{X}^N : d_{TV}\left(\Pi(\theta|\mathbf{X}^N), N\left(\hat{\theta}_N, \frac{1}{N} I_{\theta_0}^{-1}\right)\right) \geq M\frac{1}{\sqrt{N}}\right) \leq \delta. \tag{57}$$

Unfortunately, BvM results like (56) and (57) do not fully generalize to infinite dimensional spaces; see counterexamples in [9]. Regarding the asymptotic frequentist properties of posterior distributions in nonparametric models, various positive results have been obtained recently; see, e.g., [10, 26, 16, 17, 7, 8]. For the convergence rate in the nonparametric case, we refer to [10, 26, 7].

In the remainder of the section, we prove a BvM type result for the posterior distribution $\mu_N^{\boldsymbol{\eta}}$ defined by (51) and (52). If we view the observational noise η and η_i appearing in the

data as random variables, then the posterior measures appearing become random probability measures. Furthermore, exploiting the randomness of the η_i, we claim that the posterior distribution in the large data case can be rewritten in the form of the small noise case. Indeed, by completing the square, we can write the expression (52) as

$$\mu_N^{\boldsymbol{\eta}}(dx) = \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}\left|G(x^\dagger) - G(x) + \frac{1}{N}\sum_{i=1}^N \eta_i\right|^2\right) dx. \tag{58}$$

Observe that $\mathcal{L}(\frac{1}{N}\sum_{i=1}^N \eta_i) = \mathcal{L}(\frac{1}{\sqrt{N}}\eta) = N(0, \frac{1}{N}\mathbf{I}_d)$ due to the normality assumptions on η and η_i. As a consequence it makes no difference which formulation is chosen when one is concerned with the statistical dependence of $\mu_N^{\boldsymbol{\eta}}$ on the law of $\boldsymbol{\eta}$. For this reason, we will only prove the BvM result for $\mu_N^{\boldsymbol{\eta}}$ given directly in the form (51).

For notational simplicity, we write the noise level $\sqrt{\varepsilon}$ in place of $\frac{1}{\sqrt{N}}$ and consider random observations $\{y_\varepsilon\}$, generated from a truth $x^\dagger$ and normal noise η, i.e.,

$$y_\varepsilon = G(x^\dagger) + \sqrt{\varepsilon}\eta.$$

Given the same prior defined as before, we obtain the posterior distribution

$$\begin{aligned}\mu_\varepsilon^\eta(dx) &= \frac{1}{Z_{\mu,\varepsilon}^\eta} \exp\left(-\frac{1}{2\varepsilon}|y_\varepsilon - G(x)|^2 - V_0(x)\right) dx \\ &= \frac{1}{Z_{\mu,\varepsilon}^\eta} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right) dx.\end{aligned}$$

For any fixed η, let ν_ε^η be the best Gaussian measure which minimizes the Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu||\mu_\varepsilon^\eta)$ over $\mathcal{A}$. For ease of calculations, from now on we only consider the rate of convergence under Assumption 5.1(i); the other case can be dealt with in the same manner (see Remark 5.8). The main result is as follows.

Theorem 5.4. *There exists $C > 0$ such that*

$$\mathbb{E}^\eta D_{KL}(\nu_\varepsilon^\eta||\mu_\varepsilon^\eta) \le C\varepsilon \tag{59}$$

as $\varepsilon \downarrow 0$.

With the help of Pinsker's inequality (2) as well as the Markov inequality, one can derive the following BvM-type result from Theorem 5.4.

Corollary 5.5. *For any $\delta > 0$, there exists a constant $M = M(\delta) > 0$ such that*

$$\mathbb{P}^\eta\left(\eta : d_{TV}(\mu_\varepsilon^\eta, \nu_\varepsilon^\eta) \ge M\sqrt{\varepsilon}\right) \le \delta \tag{60}$$

when $\varepsilon \downarrow 0$.

Remark 5.6.

(i) Because of the statistical equivalence of posterior measures in the limit of large data size and small noise, the posterior measure $\mu_N^{\boldsymbol{\eta}}$ in the large data case (given by (52)) has the same convergence rate as (60), namely, for any $\delta > 0$, there exists a constant $M = M(\delta) > 0$ such that

$$\mathbb{P}^{\boldsymbol{\eta}}\left(\boldsymbol{\eta} : d_{\mathrm{TV}}(\mu_N^{\boldsymbol{\eta}}, \nu_N^{\boldsymbol{\eta}}) \geq M/\sqrt{N}\right) \leq \delta \tag{61}$$

as $N \to \infty$. This recovers the optimal rate of convergence for the posterior as proved for statistical models; see (57).

(ii) For a fixed realization of the noise η, we have shown in Theorem 5.2(i) that $D_{\mathrm{KL}}(\nu_N||\mu_N^{\boldsymbol{\eta}}) \downarrow 0$ as $N \to \infty$. In fact, by following the proof of the Laplace method, one can prove that $D_{\mathrm{KL}}(\nu_N||\mu_N^{\boldsymbol{\eta}}) = \mathcal{O}(1/\sqrt{N})$. However, we obtain the linear convergence rate in (59) (with ε replacing $1/N$) by utilizing the symmetric cancellations in the evaluation of Gaussian integrals.

To prove Theorem 5.4, we start with an averaging estimate for the logarithm of the normalization constant $Z_{\mu,\varepsilon}^{\eta}$.

Lemma 5.7.

$$\mathbb{E}^{\eta} \log Z_{\mu,\varepsilon}^{\eta} \leq \frac{d}{2}\log(2\pi\varepsilon) - V_0(x^\dagger) + \log\det DG(x^\dagger) + r_\varepsilon, \tag{62}$$

where $r_\varepsilon \leq C\varepsilon$ for some $C > 0$ independent of ε.

Proof. Take a constant $\gamma \in (0, \frac{1}{2})$. We write $\mathbb{E}^{\eta} \log Z_{\mu,\varepsilon}^{\eta}$ as the sum

$$\mathbb{E}^{\eta} \log Z_{\mu,\varepsilon}^{\eta} = \mathbb{E}^{\eta}\left(\log Z_{\mu,\varepsilon}^{\eta}\mathbf{1}_{|\eta|\leq\varepsilon^{-\gamma}}\right) + \mathbb{E}^{\eta}\left(\log Z_{\mu,\varepsilon}^{\eta}\mathbf{1}_{|\eta|\geq\varepsilon^{-\gamma}}\right) =: I_1 + I_2.$$

We first find an upper bound for I_2. By definition,

$$\begin{aligned} Z_{\mu,\varepsilon}^{\eta} &= \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right) dx \\ &\leq \int_{\mathbb{R}^d} e^{-V_0(x)} dx = Z_0. \end{aligned}$$

It follows that

$$I_2 \leq \log Z_0 \cdot P^{\eta}(\eta : |\eta| \geq \varepsilon^{-\gamma}) \leq \log Z_0 \cdot e^{-\varepsilon^{-2\gamma}}.$$

For I_1, we need to estimate $Z_{\mu,\varepsilon}^{\eta}$ under the assumption that $|\eta| \leq \varepsilon^{-\gamma}$. Thanks to condition (i) on G, when $\varepsilon \ll 1$ there exists a unique $m_{\varepsilon,\eta}^\dagger$ such that $G(m_{\varepsilon,\eta}^\dagger) = G(x^\dagger) + \sqrt{\varepsilon}\eta$. Moreover, denoting by H the inverse of G in the neighborhood of $G(x^\dagger)$, we get from Taylor expansion that

$$m_{\varepsilon,\eta}^\dagger = x^\dagger + DH(G(x^\dagger))\sqrt{\varepsilon}\eta + \varepsilon \sum_{|\alpha|=2} \partial_\alpha H(\xi G(x^\dagger) + (1-\xi)\sqrt{\varepsilon}\eta)\eta^\alpha, \tag{63}$$

with some $\xi \in (0,1)$. Thanks to the smoothness assumption on G, the function H is differentiable up to the fourth order and hence the coefficients in the summation are uniformly bounded. Moreover, noting that $DH(G(x^\dagger)) = DG(x^\dagger)^{-1}$, we obtain

$$m_{\varepsilon,\eta}^\dagger = x^\dagger + DG(x^\dagger)^{-1}\sqrt{\varepsilon}\eta + \varepsilon R_\varepsilon(\eta), \tag{64}$$

where $\limsup_{\varepsilon\downarrow 0}|R_\varepsilon(\eta)| \le C|\eta|^2$ for some positive C which is independent of ε and η. Next, according to the proof of Lemma 3.2, given any sufficiently small $\delta > 0$, we can write $Z^\eta_{\mu,\varepsilon} = I^{\delta,\eta}_\varepsilon + J^{\delta,\eta}_\varepsilon$, where $|J^{\delta,\eta}_\varepsilon| \le Ce^{-\frac{C}{\varepsilon}}$ with some $C > 0$ independent of η and

$$I^{\delta,\eta}_\varepsilon = \int_{B^{\delta,\eta}_\varepsilon} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right) dx$$

with $B^{\delta,\eta}_\varepsilon := B(m^\dagger_{\varepsilon,\eta}, \delta)$. Now we seek bounds for $I^{\delta,\eta}_\varepsilon$. Thanks to Assumption 5.1(i) and the fact that $m^\dagger_{\varepsilon,\eta} \to x^\dagger$, G is a C^1-diffeomorphism in the neighborhood of $m^\dagger_{\varepsilon,\eta}$. Therefore there exist positive constants $\delta_1 < \delta_2$ depending only on δ such that $B(G(m^\dagger_{\varepsilon,\eta}), \delta_1) \subset G(B^{\delta,\eta}_\varepsilon) \subset B(G(m^\dagger_{\varepsilon,\eta}), \delta_2)$. After applying the transformation $x \mapsto H(x)$ in evaluation of the integral $I^{\delta,\eta}_\varepsilon$, we get

$$\tilde{I}^{\delta_1,\eta}_\varepsilon \le I^{\delta,\eta}_{i,\varepsilon} \le \tilde{I}^{\delta_2,\eta}_\varepsilon,$$

where

$$\tilde{I}^{\delta,\eta}_\varepsilon := \int_{B(0,\delta)} \exp\left(-\frac{1}{2\varepsilon}|y|^2 - V_0 \circ H\big(y + G\big(m^\dagger_{\varepsilon,\eta}\big)\big)\right) \det\big(DH\big(y + G\big(m^\dagger_{\varepsilon,\eta}\big)\big)dy.$$

In order to estimate $\tilde{I}^{\delta,\eta}_\varepsilon$, in $B(0,\delta)$ with some small δ we define two auxiliary functions by setting

$$f_{\varepsilon,\eta}(\cdot) := \exp(-V_0 \circ H(\cdot + G(m^\dagger_{\varepsilon,\eta}))) \det(DH(\cdot + G(m^\dagger_{\varepsilon,\eta}))$$

and

$$L(\cdot) := \exp(-V_0 \circ H(G(\cdot))) \det(DH(G(\cdot)) = \exp(-V_0(\cdot))/\det(DG(\cdot)).$$

It is worthy to note that within the ball $B(0,\delta)$, all derivatives up to second order of $f_{\varepsilon,\eta}$ as well as of L can be bounded uniformly with respect to sufficiently small ε and η such that $|\eta| \le \varepsilon^{-\gamma}$. Taking (64) into account, we can expand L near $m^\dagger$ to get that

$$\begin{aligned} f_{\varepsilon,\eta}(0) &= L\big(m^\dagger_{\varepsilon,\eta}\big) = L(x^\dagger) + \nabla L(x^\dagger)^T\big(m^\dagger_{\varepsilon,\eta} - x^\dagger\big) \\ &\quad + \frac{1}{2}\big(m^\dagger_{\varepsilon,\eta} - x^\dagger\big)^T \nabla^2 L\big(\theta x^\dagger + (1-\theta)m^\dagger_{\varepsilon,\eta}\big)\big(m^\dagger_{\varepsilon,\eta} - x^\dagger\big) \\ &= \frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \varepsilon^{\frac{1}{2}} \nabla L(x^\dagger)^T DG(x^\dagger)^{-1}\eta + r_{1,\varepsilon,\eta} \end{aligned} \tag{65}$$

with some $\theta \in (0,1)$ and the residual $|r_{1,\varepsilon,\eta}| \le C\varepsilon|\eta|^2$ for some $C > 0$. Moreover, for any $y \in B(0,\delta)$,

$$f_{\varepsilon,\eta}(y) = f_{\varepsilon,\eta}(0) + \nabla f_{\varepsilon,\eta}(0)^T y + \frac{1}{2} y^T \nabla^2 f_{\varepsilon,\eta}(\xi y) y \tag{66}$$

for some $\xi = \xi(y) \in (0,1)$. Then it follows from (65) and (66) that

$$\begin{aligned} \tilde{I}^{\delta,\eta}_\varepsilon &= \int_{B(0,\delta)} \exp\left(-\frac{1}{2\varepsilon}|y|^2\right) f_{\varepsilon,\eta}(y)dy \\ &= \varepsilon^{\frac{d}{2}} \int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp\left(-\frac{1}{2}|y|^2\right) f_{\varepsilon,\eta}(\varepsilon^{\frac{1}{2}} y)dy \end{aligned} \tag{67}$$

$$= \varepsilon^{\frac{d}{2}} \left(f_{\varepsilon,\eta}(0) \int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp\left(-\frac{1}{2}|y|^2\right) dy \right.$$
$$\left. + \frac{\varepsilon}{2} \int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp\left(-\frac{1}{2}|y|^2\right) y^T \nabla^2 f_{\varepsilon,\eta}(\xi y) y dy \right)$$
$$= (2\pi\varepsilon)^{\frac{d}{2}} \left(\frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \nabla L(x^\dagger)^T DG(x^\dagger)^{-1} \sqrt{\varepsilon}\eta + r_{2,\varepsilon,\eta} \right)$$

with $|r_{2,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. Notice that the linear term in the expansion (66) vanishes from the second line to the third line because the region of integration is symmetric with respect to the origin; the final equality holds because we have counted the exponentially decaying Gaussian integral outside of the ball $B(0, \varepsilon^{-\frac{1}{2}}\delta)$ in the residual $r_{2,\varepsilon,\eta}$. Hence we obtain that for $|\eta| \leq \varepsilon^{-\gamma}$ and ε small enough

$$I^\delta_{\varepsilon,\eta} = (2\pi\varepsilon)^{\frac{d}{2}} \left(\frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \varepsilon^{\frac{1}{2}} \nabla L(x^\dagger)^T DG(x^\dagger)^{-1}\eta + r_{2,\varepsilon,\eta} \right)$$

with $|r_{2,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. As a result, $Z^\eta_{\mu,\varepsilon}$ satisfies the same bound as above. Then by using the Taylor expansion of the log function, one obtains that

$$\log Z^\eta_{\mu,\varepsilon} = \log \left(\frac{(2\pi\varepsilon)^{\frac{d}{2}} \exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} \right) + \varepsilon^{\frac{1}{2}} p^T \eta + r_{3,\varepsilon,\eta},$$

where p is vector depending only on L, G, V_0 and $x^\dagger$ and $|r_{3,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. This implies that when ε is sufficiently small,

$$I_1 = \mathbb{E}^\eta \left(\log Z^\eta_{\mu,\varepsilon} \mathbf{1}_{|\eta| \leq \varepsilon^{-\gamma}} \right) = \frac{d}{2} \log(2\pi\varepsilon) - V_0(x^\dagger) + \log\det DG(x^\dagger) + r_\varepsilon$$

with $|r_\varepsilon| \leq C\varepsilon$. Again the first order term $\varepsilon^{\frac{1}{2}} p^T \eta$ vanishes because of the symmetry of the integration region; the bound $|r_\varepsilon| \leq C\varepsilon$ follows from the bound for $r_{3,\varepsilon,\eta}$ and the Gaussian tail bound. This completes the proof. ∎

Proof of Theorem 5.4. We prove the theorem by constructing a family of Gaussian measures $\{\overline{\nu}^\eta_\varepsilon\}$ such that

$$\mathbb{E}^\eta D_{\mathrm{KL}}(\overline{\nu}^\eta_\varepsilon || \mu^\eta_\varepsilon) \leq C\varepsilon \tag{68}$$

for some $C > 0$. Then the theorem is proved by the optimality of $\nu_{\varepsilon,\eta}$. Recall that $m^\dagger_{\varepsilon,\eta}$ is defined by (63). Fixing $\gamma \in (0, \frac{1}{2})$, we define $\overline{\nu}^\eta_\varepsilon = N(\overline{m}_{\varepsilon,\eta}, \overline{\mathbf{\Sigma}}_{\varepsilon,\eta})$ with $\overline{m}_{\varepsilon,\eta}$ defined by

$$\overline{m}_{\varepsilon,\eta} = \begin{cases} m^\dagger_{\varepsilon,\eta} & \text{if } |\eta| \leq \varepsilon^{-\gamma}, \\ x^\dagger & \text{otherwise} \end{cases}$$

and $\overline{\mathbf{\Sigma}}_{\varepsilon,\eta} = \left(DG(\overline{m}_{\varepsilon,\eta})^T DG(\overline{m}_{\varepsilon,\eta}) \right)^{-1}$. Clearly, when ε is small enough, $\overline{m}_{\varepsilon,\eta}$ admits an expansion similar to (63). As a consequence, there exist positive constants C_1, c_2, C_2 which

are independent of η such that $\limsup_{\varepsilon\downarrow 0}|\overline{m}_{\varepsilon,\eta}| \le C_1$ and $c_2 \le \liminf_{\varepsilon\downarrow 0}\lambda_{\min}(\overline{\mathbf{\Sigma}}_{\varepsilon,\eta}) < \limsup_{\varepsilon\downarrow 0}\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) \le C_2$ hold for all η. With the above choice for $(\overline{m}_{\varepsilon,\eta}, \overline{\mathbf{\Sigma}}_{\varepsilon,\eta})$, an application of Lemma 3.7 with $V_1^\varepsilon(x) = \frac{1}{2}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2$ and $V_2(x) = V_0(x)$ yields that

$$D_{\mathrm{KL}}(\overline{\nu}_\varepsilon^\eta||\overline{\mu}_\varepsilon^\eta) = V_0(\overline{m}_{\varepsilon,\eta}) - \frac{d}{2}\log(2\pi\varepsilon) + \log\det DG(\overline{m}_{\varepsilon,\eta}) + \log Z_{\mu,\varepsilon}^\eta + r_\varepsilon, \tag{69}$$

where $r_\varepsilon \le C\varepsilon$ with $C = C(C_1, c_2, C_2, M_V)$. By the definition of $\overline{m}_{\varepsilon,\eta}$ and the expansion (63), it follows from the Taylor expansion for the function $x \mapsto V_0(x) + \frac{1}{2}\log\det DG(x)$ that when $|\eta| \le \varepsilon^{-\gamma}$ and ε is small enough,

$$V_0(\overline{m}_{\varepsilon,\eta}) + \log\det DG(\overline{m}_{\varepsilon,\eta}) = V_0(x^\dagger) + \log\det DG(x^\dagger) + \sqrt{\varepsilon}q^T\eta + \tilde{r}_{\varepsilon,\eta} \tag{70}$$

with some $q \in \mathbb{R}^d$ and $|\tilde{r}_{\varepsilon,\eta}| \le C\varepsilon$ for some $C > 0$. Then the estimate (68) follows, by taking the expectation of (69) and using (70) and Lemma 3.7. ∎

Remark 5.8. Theorem 5.4 proves the rate of convergence with the assumption that G satisfies Assumption 5.1(i). However, the convergence rate remains the same when Assumption 5.1(ii) is fulfilled and when the best Gaussian measure is replaced by the best Gaussian mixture.

5.3. Comparison with classical BvM results. We would like to make comparisons between our BvM result for Bayesian inverse problems and classical finite dimensional BvM results for general statistical models [11, 13].

- *Assumption.* In the classical framework of Bayesian inferences, the posterior converges to a Gaussian in the total variation distance (with optimal rate) under the typical assumption that the likelihood function is C^3 and that the Fisher information matrix is nondegenerate; see, e.g., [11, Theorem 1.4.2] and [13, section 4]. The asymptotic covariance of the limiting Gaussian is given by the inverse of the Fisher information matrix. In the Bayesian inverse problem setting, we improve the convergence to the stronger sense of Kullback–Leibler divergence, but at the expense of requiring higher differentiability (C^4) on the forward map G. Moreover, the matrix product DG^TDG takes the place of the Fisher information matrix in the asymptotic covariance, where DG is invertible because of Assumption 5.1.
- *Multimodal distribution.* The proposed Kullback–Leibler approximation framework allows us to prove the convergence of a multimodal probability measure to a mixture of Gaussian measures. The limiting Kullback–Leibler discrepency between the target measure and the Gaussian approximation is characterized explicitly as a sum of two relative entropies; see Theorem 4.2. In addition, in this case the prior does not disappear in the limit and its influence on the posterior is reflected in the weighted coefficients in the Gaussian mixture. To the best of our knowledge, such results have not been stated in the statistical literature.
- *Proof.* Both our proof and classical proofs for the finite dimensional BvM theorems are essentially based on the local Taylor expansion of the posterior around the truth. But the proofs are carried out in different ways. Classical BvM results in the total variation distance are usually proved by first expanding the posterior density around

the MLE, which requires tracking the normalization constant, and then applying the local asymptotic normality of MLE and LeCam's contiguity arguments to obtain the convergence of the posterior. Our proof, instead, takes advantage of the special formulation of the Kullback–Leibler divergence, i.e., the separation of the normalization constant from the log density, thereby reducing the convergence proof to establishing precise estimates on the normalization constant (see Lemma 5.7).

6. Conclusions. We have studied a methodology widely used in applications, yet little analyzed, namely, the approximation of a given target measure by a Gaussian, or by a Gaussian mixture. We have employed relative entropy as a measure of goodness of fit. Our theoretical framework demonstrates the existence of minimizers of the variational problem and studies their asymptotic form in a relevant small parameter limit where the measure concentrates; the small parameter limit is studied by use of tools from Γ-convergence. In the case of a target with asymptotically unimodal distribution the Γ-limit demonstrates perfect reconstruction by the approximate single Gaussian method in the measure concentration limit, and in the case of multiple modes it quantifies the errors resulting from using a single mode fit. Furthermore the Gaussian mixture is shown to overcome the limitations of a single mode fit, in the case of target measure with multiple modes. These ideas are exemplified in the analysis of a Bayesian inverse problem in the small noise or large data set limits and connections made to the BvM theory from asymptotic statistics.

The BvM theorem of this paper is essentially still parametric. A natural interesting future direction would be to study infinite dimensional statistical models [12]. In particular it would be interesting to apply our measure approximation approach from Γ-convergence to understand the BvM phenomenon of infinite dimensional nonlinear Bayesian inverse problems. In our finite dimensional setting, the inverse problem of interest is essentially well-posed since we assume that both G and DG are invertible, so the only ill-posedness comes from the lack of uniqueness. However, for infinite dimensional inverse problems, the degree of ill-posedness (mild/severe) has a big influence on the precise statement of the BvM theorem. Understanding this issue requires delicate quantitative stability estimates for the underlying inverse problem. The recent paper [18] proved a BvM result for high dimensional nonlinear inverse problems where the dimension of the unknown parameter increases with the decreasing noise level. However, it remains an open problem whether the BvM theorem holds for genuinely infinite dimensional nonlinear inverse problems. We will address this problem in future work.

REFERENCES

[1] S. Agapiou, S. Larsson, and A. Stuart, *Posterior consistency of the Bayesian approach to linear ill-posed inverse problems*, Stochastic Process. Appl., 123 (2013), pp. 3828–3860.
[2] A. Barron, M. J. Schervish, and L. Wasserman, *The consistency of posterior distributions in nonparametric problems*, Ann. Statist., 27 (1999), pp. 536–561.
[3] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
[4] A. Braides, Γ-*Convergence for Beginners*, Oxford University Press, Oxford, UK, 2002.
[5] L. D. Brown and M. G. Low, *Asymptotic equivalence of nonparametric regression and white noise*, Ann. Statist., 24 (1996), pp. 2384–2398.
[6] L. M. L. Cam, *Convergence of estimates under dimensionality restrictions*, Ann. Statist., 1 (1973), pp. 38–53.

[7] I. Castillo and R. Nickl, *Nonparametric Bernstein-von Mises theorems in Gaussian white noise*, Ann. Statist., 41 (2013), pp. 1999–2028.
[8] I. Castillo and R. Nickl, *On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures*, Ann. Statist., 42 (2014), pp. 1941–1969.
[9] D. Freedman, *On the Bernstein-von Mises theorem with infinite-dimensional parameters*, Ann. Statist., 27 (1999), pp. 1119–1141.
[10] S. Ghosal, J. Ghosal, and A. W. van der Vaart, *Convergence rates of posterior distributions*, Ann. Statist., 28 (2000), pp. 500–531.
[11] J. K. Ghosh and R. V. Ramamoorthi, *Bayesian Nonparametrics*, Springer Ser. Statist., Springer, New York, 2003.
[12] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Camb. Ser. Stat. Probab. Math. 40, Cambridge University Press, Cambridge, UK, 2015.
[13] C. Hipp and R. Michel, *On the Bernstein-von Mises approximation of posterior distributions*, Ann. Statist., 4 (1976), pp. 972–980.
[14] J. L. Jensen, *Saddlepoint Approximations*, Oxford Statist. Sci. Ser. 16, Oxford University Press, Oxford, UK, 1995.
[15] M. Katsoulakis and P. Plecháč, *Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems*, J. Chem, Phys, 139 (2013), 074115.
[16] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten, *Bayesian inverse problems with Gaussian priors*, Ann. Statist., 39 (2011), pp. 2626–2657.
[17] H. Leahu, *On the Bernstein-von Mises phenomenon in the Gaussian white noise model*, Electron. J. Statist., 5 (2011), pp. 373–404.
[18] Y. Lu, *On the Bernstein-von Mises Theorem for High Dimensional Nonlinear Bayesian Inverse Problems*, preprint, arXiv:1706.00289, 2017.
[19] Y. Lu, A. Stuart, and H. Weber, *Gaussian approximations for transition paths in Brownian dynamics*, SIAM J. Math. Anal., 49 (2017), pp. 3005–3047.
[20] A. J. Majda and B. Gershgorin, *Improving model fidelity and sensitivity for complex systems through empirical information theory*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 10044–10049.
[21] G. Menz and A. Schlichting, *Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape*, Ann. Probab., 42 (2014), pp. 1809–1884.
[22] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber., *Algorithms for Kullback-Leibler approximation of probability measures in infinite dimensions*, SIAM J. Sci. Comput., 37 (2015), pp. A2733–A2757.
[23] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber., *Kullback-Leibler approximation for probability measures in infinite dimensional spaces*, SIAM J. Math. Anal, 47 (2015), pp. 4091–4122.
[24] D. Sanz-Alonso and A. M. Stuart, *Gaussian approximations of small noise diffusions in Kullback–Leibler divergence*, Commun. Math. Sci., 15 (2017), pp. 2087–2097.
[25] L. Schwartz, *On Bayes procedures*, Probab. Theory and Related Fields, 4 (1965).
[26] X. Shen and L. Wasserman, *Rates of convergence of posterior distributions*, Ann. Statist., 29 (2001), pp. 687–714.
[27] A. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[28] A. Van Der Vaart, *Asymptotic Statistics*, Vol. 3, Cambridge University Press, Cambridge, UK, 2000.
[29] S. J. Vollmer, *Posterior consistency for Bayesian inverse problems through stability and regression results*, Inverse Problems, 29 (2013), 125011.
[30] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Machine Learning, 1 (2008), pp. 1–305.